Supplementary Materials for Balanced and Robust Randomized Treatment Assignments: The Finite Selection Model for the Health Insurance Experiment and Beyond*

Ambarish Chattopadhyay†, Carl N. Morris‡, José R. Zubizarreta§

*This work was supported through a grant from the Alfred P. Sloan Foundation (G-2020-13946).

†Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata, West Bengal, 700108; email: ambarish@isical.ac.in.

‡Department of Statistics, Harvard University, 1 Oxford Street Cambridge, MA 02138; email: morris@stat.harvard.edu.

§Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University, 180 Longwood Avenue, Office 307-D, Boston, MA 02115; email: zubizarreta@hcp.med.harvard.edu.

A Notation, estimands, and acronyms

Table A1: Notation

N	$\triangleq$	Full sample size
i	$\triangleq$	Index of unit, $i = 1, ..., N$
G	$\triangleq$	Number of treatments
g	$\triangleq$	Index of treatment group, $g = 1, 2, ..., G$
n_g	$\triangleq$	Size of treatment group g
k	$\triangleq$	Number of baseline covariates
$\boldsymbol{X}_i$	$\triangleq$	Observed vector of baseline covariates of unit i
$\underline{\boldsymbol{X}}_{\text{full}}$	$\triangleq$	$N \times k$ matrix of covariates in the full sample
$\underline{\tilde{\boldsymbol{X}}}_{\text{full}}$	$\triangleq$	$N \times k + 1$ design matrix in the full sample
$\bar{\boldsymbol{X}}_{\text{full}}$	$\triangleq$	$k \times 1$ vector of means of the baseline covariates in the full sample
$\underline{\boldsymbol{S}}_{\text{full}}$	$\triangleq$	$k \times k$ covariance matrix of the baseline covariates in the full sample
$Y_i(g)$	$\triangleq$	Potential outcome of unit i under treatment g
$\boldsymbol{Y}(g)$	$\triangleq$	Vector of potential outcomes under treatment g, $(Y_1(g), ..., Y_N(g))^\top$
Z_i	$\triangleq$	Treatment assignment indicator of unit i, $Z_i \in \{1, 2, ..., G\}$
$\boldsymbol{Z}$	$\triangleq$	Vector of treatment assignment indicators, $(Z_1, ..., Z_N)^\top$
Y_i^{obs}	$\triangleq$	Observed outcome of unit i, $Y_i^{\text{obs}} = \sum_{g=1}^{G} \mathbb{1}(Z_i = g) Y_i(g)$

Table A2: Estimands

$Y_i(g') - Y_i(g'')$	$\triangleq$	Unit level causal effect of treatment g' relative to treatment g'' for unit i; $g', g'' \in \{1, 2, ..., G\}$
$\text{SATE}_{g',g''}$	$\triangleq$	$\frac{1}{N} \sum_{i=1}^{N} \{Y_i(g') - Y_i(g'')\}$, the Sample Average Treatment Effect of treatment g' relative to treatment g''
$\text{PATE}_{g',g''}$	$\triangleq$	$\mathbb{E}\{Y_i(g') - Y_i(g'')\}$, the Population Average Treatment Effect of treatment g' relative to treatment g''

Table A3: Acronyms

ASMD	Absolute Standardized Mean Difference
CRD	Completely Randomized Design
FSM	Finite Selection Model
HIE	Health Insurance Experiment
OLS	Ordinary Least Squares
PATE	Population Average Treatment Effect
RBD	Randomized Block Design
RR	Re-Randomization
SATE	Sample Average Treatment Effect
SCOMARS	Sequentially Controlled Markovian Random Sampling
SOM	Selection Order Matrix

B Proofs of theoretical results

Lemma A1. Let treatment 1 be the choosing group at the rth stage. Also, let $\underline{\tilde{\boldsymbol{X}}}_{r-1}$ be the $\tilde{n}_{r-1} \times (k+1)$ design matrix in treatment group 1 after the $(r-1)$th stage, where $\tilde{n}_{r-1} \geq 1$ and $\text{rank}(\underline{\tilde{\boldsymbol{X}}}_{r-1}) = k+1$. The D-optimal selection function chooses unit i' with covariate vector $\boldsymbol{X}_{i'} \in \mathbb{R}^k$, where

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \tag{A1}$$

Proof. We follow the notations outlined in Section 4. At the rth stage, D-optimal selection function selects unit $i' \in \mathcal{R}_{r-1}$, where $i' \in \arg\max_{i \in \mathcal{R}_{r-1}} \det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i})$. Now, for $i \in \mathcal{R}_{r-1}$,

$$\det(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i}) = \det\left\{ \underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} (1, \boldsymbol{X}_i^\top) \right\} \tag{A2}$$

$$= \det(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}) \det\left\{ \underline{\boldsymbol{I}} + (\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-\frac{1}{2}} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} (1, \boldsymbol{X}_i^\top) (\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-\frac{1}{2}} \right\} \tag{A3}$$

$$= \det(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}) \Big\{ 1 + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \Big\}, \tag{A4}$$

where the final equality holds since for two matrices $\underline{\boldsymbol{A}}_{m\times n}$ and $\underline{\boldsymbol{B}}_{n\times m}$, $\det(\underline{\boldsymbol{I}}_m + \underline{\boldsymbol{AB}}) = \det(\underline{\boldsymbol{I}}_n + \underline{\boldsymbol{BA}})$. Equation A4 implies that the selected unit i' maximizes $(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}$. This completes the proof.

□

Proof of Theorem 4.1

Proof. We use the notations in Section 3.1 and Table A1. We first consider the case where $\tilde{n}_{r-1} = 0$. The selected unit i' satisfies,

$$i' \in \arg\max_{i\in\mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}. \tag{A5}$$

Now, denoting $\boldsymbol{e}_1 = (1, 0, ..., 0)$ as the $k \times 1$ first standard unit vector, we have

$$\begin{aligned}
(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} &= (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}} \frac{\boldsymbol{e}_1}{N} + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= \frac{1}{N} + \{0, (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top\}(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}} \end{pmatrix} \\
&= \frac{1}{N} + \frac{1}{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}).
\end{aligned} \tag{A6}$$

Here the last equality holds since, by the formula for the inverse of a partitioned matrix, $(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} = \begin{pmatrix} \underline{\boldsymbol{B}}_{11} & \underline{\boldsymbol{B}}_{12} \\ \underline{\boldsymbol{B}}_{21} & \underline{\boldsymbol{B}}_{22} \end{pmatrix}$, where $\underline{\boldsymbol{B}}_{22}^{-1} = \underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - N\bar{\boldsymbol{X}}_{\text{full}} \bar{\boldsymbol{X}}_{\text{full}}^\top = N\underline{\boldsymbol{S}}_{\text{full}}$. This completes the proof of the $\tilde{n}_{r-1} = 0$ case. The proof for the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible follows similar steps and hence is omitted.

We now consider the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1}$ is not invertible. We denote $\bar{\boldsymbol{X}}_{r-1}^* = \frac{\bar{\boldsymbol{X}}_{r-1} + \epsilon \bar{\boldsymbol{X}}_{\text{full}}}{1+\epsilon}$ and $\underline{\boldsymbol{S}}_{r-1}^* = (\frac{1}{\tilde{n}_{r-1}} \underline{\boldsymbol{X}}_{r-1}^\top \underline{\boldsymbol{X}}_{r-1} + \frac{\epsilon}{N} \underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}}) - (1+\epsilon)\bar{\boldsymbol{X}}_{r-1}^* \bar{\boldsymbol{X}}_{r-1}^{*\top}$. The selected unit i' satisfies,

$$i' \in \arg\max_{i\in\mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top) \Big(\frac{1}{\tilde{n}_{r-1}} \underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N} \underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}} \Big)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \tag{A7}$$

Denoting $\tilde{\underline{\boldsymbol{G}}} = \begin{pmatrix} \sqrt{\frac{1}{\tilde{n}_{r-1}}}\underline{\boldsymbol{X}}_{r-1} \\ \sqrt{\frac{\epsilon}{N}}\underline{\boldsymbol{X}}_{\text{full}} \end{pmatrix}$, we have

$$
\begin{aligned}
&(1, \boldsymbol{X}_i^\top)\Big(\frac{1}{\tilde{n}_{r-1}}\tilde{\underline{\boldsymbol{X}}}_{r-1}^\top\tilde{\underline{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N}\tilde{\underline{\boldsymbol{X}}}_{\text{full}}^\top\tilde{\underline{\boldsymbol{X}}}_{\text{full}}\Big)^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \\
&= (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^* \end{pmatrix} + (1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix} 1 \\ \bar{\boldsymbol{X}}_{r-1}^* \end{pmatrix} \\
&= (0, (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*)^\top)(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1}\begin{pmatrix} 0 \\ \boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^* \end{pmatrix} + \frac{1}{1+\epsilon} \\
&= (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*)^\top(\underline{\boldsymbol{S}}_{r-1}^*)^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}^*) + \frac{1}{1+\epsilon}. \qquad \text{(A8)}
\end{aligned}
$$

Here, the third equality holds since $\begin{pmatrix} 1 \\ \bar{\boldsymbol{X}}_{r-1}^* \end{pmatrix} = \frac{1}{1+\epsilon}\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}}\boldsymbol{e}_1$ and the fourth equality holds since $(\tilde{\underline{\boldsymbol{G}}}^\top\tilde{\underline{\boldsymbol{G}}})^{-1} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \underline{\boldsymbol{B}}_{21} & \underline{\boldsymbol{B}}_{22} \end{pmatrix}$, where $\underline{\boldsymbol{B}}_{22}^{-1} = (\frac{1}{\tilde{n}_{r-1}}\underline{\boldsymbol{X}}_{r-1}^\top\underline{\boldsymbol{X}}_{r-1} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top\underline{\boldsymbol{X}}_{\text{full}}) - (1+\epsilon)\bar{\boldsymbol{X}}_{r-1}^*\bar{\boldsymbol{X}}_{r-1}^{*\top} = \underline{\boldsymbol{S}}_{r-1}^*$. This completes the proof. □

Proof of Theorem 4.2

Proof. (a) We first consider the setting of a standard block design where $N = BG$ (i.e., $c = 1$). The blocks are labelled $1, 2, ..., B$. Here, the SOM is constructed by stacking B independent random permutations of the 'chunk' $(1, 2, ..., G)$. We will show that the choices made by the treatment groups in the FSM follow the assignment mechanism of an RBD.

Consider the first randomized chunk of the SOM, i.e., a random permutation of $(1, 2, ..., G)$. At the first stage of this randomized chunk, the choosing treatment group aims to maximize $(1, \boldsymbol{X}_i^\top)(\tilde{\underline{\boldsymbol{X}}}_{\text{full}}^\top\tilde{\underline{\boldsymbol{X}}}_{\text{full}})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}$. Note that we can write $\tilde{\underline{\boldsymbol{X}}}_{\text{full}}$ as $\tilde{\underline{\boldsymbol{X}}}_{\text{full}} = \begin{pmatrix} \underline{\boldsymbol{D}} \\ \vdots \\ \underline{\boldsymbol{D}} \end{pmatrix}$, where $\underline{\boldsymbol{D}}_{B\times B} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$. Now, consider a transformation of the rows of the design matrix given by $\tilde{\tilde{\boldsymbol{X}}}_i = (\underline{\boldsymbol{D}}^\top)^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}$. The transformed design matrix is $\tilde{\tilde{\underline{\boldsymbol{X}}}}_{\text{full}} = \tilde{\underline{\boldsymbol{X}}}_{\text{full}}\underline{\boldsymbol{D}}^{-1} = \begin{pmatrix} \underline{\boldsymbol{I}}_B \\ \vdots \\ \underline{\boldsymbol{I}}_B \end{pmatrix}$. We note

that the $\tilde{\tilde{\boldsymbol{X}}}_i$s nothing but standard unit vectors. Now,

$$(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} = \tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i. \tag{A9}$$

Therefore, the selection function remains the same under the above transformation. Now, $\tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i = \frac{1}{G} \tilde{\tilde{\boldsymbol{X}}}_i^\top \tilde{\tilde{\boldsymbol{X}}}_i = \frac{1}{G}$ for all i, which essentially implies that the choosing group has no preference among the units for selection and hence chooses any one of the N units randomly. Similarly, at the subsequent stages of this randomized chunk, the corresponding choosing groups select one of the remaining units randomly.

Next, we consider the second randomized chunk of the SOM. Without loss of generality, suppose treatment 1 gets to choose first in this chunk. Also, without loss of generality, suppose that in its first choice, treatment 1 had selected a unit from block 1. We claim that in this selection, treatment 1 will choose one of the remaining units randomly from any block other than block 1, which respects the assignment mechanism of an RBD.

To prove the claim, we first consider the objective function at this stage. Treatment 1 aims to maximize $(1, \boldsymbol{X}_i^\top)\Big(\frac{1}{\tilde{n}_{r-1}} \underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \frac{\epsilon}{N} \underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}}\Big)^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}$. Here, we denote the current stage by r. Using the same transformation as in the case of the first chunk, we can write the objective function as $\tilde{\tilde{\boldsymbol{X}}}_i^\top \Big(\frac{1}{\tilde{n}_{r-1}} \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} + \frac{\epsilon}{N} \underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}\Big)^{-1} \tilde{\tilde{\boldsymbol{X}}}_i$, where $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \underline{\tilde{\boldsymbol{X}}}_{r-1} \underline{\boldsymbol{D}}^{-1}$. Since $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{\text{full}} = G \underline{\boldsymbol{I}}_B$, it is equivalent to maximize

$$\tilde{\tilde{\boldsymbol{X}}}_i^\top \Big(\underline{\boldsymbol{I}}_b + \frac{B}{\tilde{n}_{r-1}\epsilon G} \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big)^{-1} \tilde{\tilde{\boldsymbol{X}}}_i = \tilde{\tilde{\boldsymbol{X}}}_i^\top \Big(\underline{\boldsymbol{I}}_b + \delta \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big)^{-1} \tilde{\tilde{\boldsymbol{X}}}_i \tag{A10}$$

$$= \tilde{\tilde{\boldsymbol{X}}}_i^\top \Big\{\underline{\boldsymbol{I}}_b - \delta \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top (\underline{\boldsymbol{I}}_{\tilde{n}_{r-1}} + \delta \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top)^{-1} \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}\Big\} \tilde{\tilde{\boldsymbol{X}}}_i. \tag{A11}$$

Here, $\delta = \frac{B}{\tilde{n}_{r-1}\epsilon G}$. The final equality holds by the Woodbury matrix identity. Now, in this case, $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = (1, 0, ..., 0)$ (since treatment 1 has only selected one unit from block 1 up to this stage). So, the objective function in Equation A11 equals $1 - \frac{\delta}{1+\delta} \tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}) \tilde{\tilde{\boldsymbol{X}}}_i^\top$. Since

$\delta > 0$, it is equivalent to minimize $\tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}) \tilde{\tilde{\boldsymbol{X}}}_i^\top = \tilde{\tilde{\boldsymbol{X}}}_i^\top \begin{pmatrix} 1 & \mathbf{0}_{1\times(B-1)}^\top \\ \mathbf{0}_{(B-1)\times 1} & \underline{\mathbf{0}}_{(B-1)\times(B-1)} \end{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i$, which takes the value 0 for a unit in any block other than block 1 and 1 for a unit in block 1. This proves the claim for treatment 1. Moreover, by similar reasoning, the claim holds for all the other treatment groups in this randomized chunk.

Next, we consider a general randomized chunk of the SOM. Once again, without loss of generality, suppose treatment 1 gets to choose first in this chunk. Also, for simplicity of exposition and without loss of generality, suppose treatment 1 has already selected from blocks $1, 2, ..., b$, implying that $\tilde{n}_{r-1} = b$ and $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = (\underline{\boldsymbol{I}}_b \, \underline{\mathbf{0}}_{b\times(B-b)})$. This form of $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}$, along with Equation A11 implies that it is equivalent to minimize $\tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}) \tilde{\tilde{\boldsymbol{X}}}_i^\top = \tilde{\tilde{\boldsymbol{X}}}_i^\top \begin{pmatrix} \underline{\boldsymbol{I}}_b & \underline{\mathbf{0}}_{b\times(B-b)} \\ \underline{\mathbf{0}}_{(B-b)\times b}^\top & \underline{\mathbf{0}}_{(B-b)\times(B-b)} \end{pmatrix} \tilde{\tilde{\boldsymbol{X}}}_i$, which is minimized for any unit i belonging to the blocks $b+1, ..., B$. This shows that at this stage, treatment 1 randomly chooses a unit from a block other than the blocks it has already chosen from. By similar reasoning, at subsequent stages of this randomized chunk, the choosing group follows the same selection strategy for their own group. This completes the proof of the theorem for the setting of a standard block design.

We now prove the theorem for the general block design setting with $N = cBG$, $c > 1$. The proof strategy is exactly the same as the $c = 1$ setting. Here the SOM is generated by randomly permuting the chunk $(1, 2, ..., G)$ $B \times c$ times. Once the selections are completed for the the first B chunks, the resulting assignment resembles that of a standard RBD (by the previous proof), where each treatment group randomly chooses exactly one unit from each block. For the $(B+1)$th chunk, suppose, without loss of generality, that treatment 1 gets to choose first. At this stage (denoted by stage r), treatment 1 tries to maximize,

$$\begin{aligned}(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} &= \tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i \\ &= \tilde{\tilde{\boldsymbol{X}}}_i^\top \tilde{\tilde{\boldsymbol{X}}}_i = 1, \end{aligned} \tag{A12}$$

where the penultimate equality holds since $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \underline{\boldsymbol{I}}_B$. Thus, similar to the first randomized

chunk in the setting of $c = 1$, treatment 1 (and the other treatments) randomly chooses one of the available units.

Finally, we consider a general chunk. Without loss of generality, suppose treatment 1 gets to choose first in this chunk. We can write the corresponding transformed design matrix $\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}$ as

$$\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1} = \begin{pmatrix} \underline{\boldsymbol{I}}_B & \\ \underline{\boldsymbol{I}}_B & \\ \vdots & \\ \underline{\boldsymbol{I}}_B & \\ \underline{\boldsymbol{I}}_b & \underline{\mathbf{0}}_{b\times(B-b)} \end{pmatrix}. \tag{A13}$$

Here, without loss of generality, we have assumed that treatment 1 has chosen $c_0 + b$ times from the first b blocks and c_0 times from the remaining blocks, where $c_0 < c$. This implies that treatment 1 aims to maximize.

$$\tilde{\tilde{\boldsymbol{X}}}_i^\top (\underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1}^\top \underline{\tilde{\tilde{\boldsymbol{X}}}}_{r-1})^{-1} \tilde{\tilde{\boldsymbol{X}}}_i = \tilde{\tilde{\boldsymbol{X}}}_i^\top \left\{ c_0 \underline{\boldsymbol{I}}_B + \begin{pmatrix} \underline{\boldsymbol{I}}_b \\ \underline{\mathbf{0}}_{(B-b)\times b}^\top \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{I}}_b & \underline{\mathbf{0}}_{b\times(B-b)} \end{pmatrix} \right\}^{-1} \tilde{\tilde{\boldsymbol{X}}}_i, \tag{A14}$$

which has the same form as the objective function in Equation A10 in the $c = 1$ setting. Thus, following similar arguments as in the $c = 1$ setting, we conclude that at this stage, treatment 1 selects a unit randomly from blocks $b+1, ..., B$, which conforms to the assignment mechanism of an RBD. Also, at subsequent stages of the randomized chunk, the choosing group follows the same selection strategy for their own group. This completes the proof of the theorem.

(b) With two groups of equal sizes, the SOM consists of successive random permutations of the 'chunk' $(1, 2)$. By Theorem 4.1, for the first pair of stages of selection, the objective function (to maximize) is given by

$$(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^\top (\underline{\boldsymbol{S}}_{\text{full}})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}). \tag{A15}$$

Under the assumption of identical twins and continuous data generating distributions, with

probability 1, there are exactly two units (one being a twin of the other), whose common covariate value $\boldsymbol{X}^{(1)}$ (say) maximizes the objective function in Equation A15. Therefore, the choosing group at the first stage selects one of these two identical twins randomly, and in the next stage, the other treatment selects the remaining twin. This respects the assignment mechanism of a matched-pair design.

Consider the next pairs of stages. The objective function of the choosing treatment group is given by:

$$(\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)})^\top \Big\{ \boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}_{\text{full}}^\top \underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} \Big\}^{-1} (\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}) \tag{A16}$$

Similar to the previous case, here also we have (with probability 1) exactly two units, one being a twin of the other, whose common covariate value $\boldsymbol{X}^{(2)}$ maximizes the objective function in Equation A16. Thus, the choosing group at the first stage of this pair selects one of these two twins randomly, and in the next stage, the other treatment chooses the remaining twin. Proceeding in this manner, it follows that, at the end of the selection process, each treatment group ends up selecting one twin randomly from $\frac{N}{2}$ identical twins, which is equivalent to a matched-pair design. This completes the proof.

□

Proof of Proposition 5.1

With equal-sized groups, by symmetry, every unit has an equal chance of belonging to one of the G treatment groups. That is, $P(Z_i = g) = \frac{1}{G}$ for all $g \in \{1, 2, ..., G\}$. Therefore,

$$\begin{aligned}\mathbb{E}\Big\{\frac{1}{n_g}\sum_{i:Z_i=g} Y_i^{\text{obs}}\Big|\boldsymbol{Y}(g)\Big\} &= \mathbb{E}\Big\{\frac{G}{N}\sum_{i=1}^{N}\mathbb{1}(Z_i = g)Y_i(g)\Big|\boldsymbol{Y}(g)\Big\} \\ &= \frac{G}{N}\sum_{i=1}^{N} P(Z_i = g)Y_i(g) \\ &= \frac{1}{N}\sum_{i=1}^{N} Y_i(g). \qquad \text{(A17)}\end{aligned}$$

Using linearity of expectations, the proposition follows from Equation A17.

Next, we derive the randomization-based variance of the estimated SATE. For simplicity, and without loss of generality, we consider the case with $G = 2$ treatment groups of equal size, and focus on the estimand $\text{SATE}_{2,1}$. Let the corresponding unbiased estimator be denoted by $\widehat{\text{SATE}}_{2,1}$. Let $W_i = \mathbb{1}(Z_i = 2)$ be the indicator that unit i belongs to group 2. Following the Neymanian decomposition in Mukerjee et al. (2018), Proposition A1 presents the closed-form expression of the variance of $\widehat{\text{SATE}}_{2,1}$.

Proposition A1.

$$\begin{aligned}&\text{Var}(\widehat{\text{SATE}}_{2,1}) \\ &= -\frac{1}{N(N-1)}\sum_{i=1}^{N}(Y_i(2) - Y_i(1) - \tau)^2 + \frac{1}{N^2}\left(\sum_{i=1}^{N} 2\{Y_i^2(1) + Y_i^2(2)\} + \right. \\ &\quad + 2\sum_{i<i'}\sum\left[Y_i(2)Y_{i'}(2)\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\} + Y_i(1)Y_{i'}(1)\left\{4\pi_{ii'}(1,1) - \frac{N}{N-1}\right\}\right] \\ &\quad \left. - 2\sum_{i<i'}\sum\left[Y_i(2)Y_{i'}(1)\left\{4\pi_{ii'}(2,1) - \frac{N}{N-1}\right\} + Y_i(1)Y_{i'}(2)\left\{4\pi_{ii'}(1,2) - \frac{N}{N-1}\right\}\right]\right),\end{aligned}$$

where $\pi_{i,i'}(z, z') = P(Z_i = z, Z_{i'} = z')$, for $z, z' \in \{1, 2\}$.

Moreover, if $\pi_{i,i'}(z, z') > 0$ for all i, i' and z, z', then a conservative estimator of this variance

is given by,

$$\widehat{\mathrm{Var}}(\widehat{\mathrm{SATE}}_{2,1}) = \frac{1}{N^2}\left(\sum_{i=1}^{N} 4(Y_i^{\mathrm{obs}})^2 \right.$$
$$+ 2\sum_{i<i'}\sum \left[\frac{W_i W_{i'} Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(2,2)}\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\}\right.$$
$$\left. + \frac{(1-W_i)(1-W_{i'})Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(1,1)}\left\{4\pi_{ii'}(1,1) - \frac{N}{N-1}\right\}\right]$$
$$- 2\sum_{i<i'}\sum \left[\frac{W_i(1-W_{i'})Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(2,1)}\left\{4\pi_{ii'}(2,1) - \frac{N}{N-1}\right\}\right.$$
$$\left.\left. + \frac{(1-W_i)W_{i'}Y_i^{\mathrm{obs}} Y_{i'}^{\mathrm{obs}}}{\pi_{ii'}(1,2)}\left\{4\pi_{ii'}(1,2) - \frac{N}{N-1}\right\}\right]\right),$$

This estimator is unbiased when treatment effect is constant across units, i.e., $Y_i(2) - Y_i(1) = c$ for all $i \in \{1, 2, ..., N\}$, where c is a constant.

When the condition $\pi_{i,i'}(z, z') > 0$ is violated for some i, i', z, z', we can still obtain a conservative variance estimator. For instance, suppose $\pi_{ii'}(1,1) = 0$. In this case, following Aronow and Samii (2013), we can upper bound the term $Y_i(2)Y_{i'}(2)\left\{4\pi_{ii'}(2,2) - \frac{N}{N-1}\right\} = -Y_i(2)Y_{i'}(2)\frac{N}{N-1}$ by $\frac{N}{2(N-1)}\{Y_i^2(2) + Y_{i'}^2(2)\}$, which admits an unbiased estimator given by $\frac{N}{N-1}W_i\{(Y_i^{\mathrm{obs}})^2 + (Y_{i'}^{\mathrm{obs}})^2\}$.

C Properties of D-optimal selection function

C.1 Affine invariance and covariate balance

Theorem A2. (a) The FSM with the D-optimal selection function is invariant under affine transformations of the covariate vector.

(b) For continuous, symmetrically distributed covariates and two groups of equal size, the FSM with the D-optimal selection function almost surely produces exact mean-balance on all even transformations of the centered covariate vector.

Proof of Theorem A2

Proof. (a) We consider the case where $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible. The proofs for the other two cases are similar. By Theorem 4.1, in this case, the chosen unit i' satisfies,

$$i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^{\top}(\underline{\boldsymbol{S}}_{r-1})^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}). \tag{A18}$$

Consider an affine transformation of the covariate $\boldsymbol{X}$ given by $\boldsymbol{U} = \underline{\boldsymbol{A}}\boldsymbol{X} + \boldsymbol{b}$, where $\underline{\boldsymbol{A}}$ is a $k \times k$ invertible matrix and $\boldsymbol{b}$ is a vector of dimension k. Let the corresponding values of $\bar{\boldsymbol{X}}_{r-1}$ and $\underline{\boldsymbol{S}}_{r-1}$ be $\bar{\boldsymbol{U}}_{r-1}$ and $\underline{\boldsymbol{S}}_{U,r-1}$, respectively. We observe that,

$$\begin{aligned}(\boldsymbol{U}_i - \bar{\boldsymbol{U}}_{r-1})^{\top}(\underline{\boldsymbol{S}}_{U,r-1})^{-1}(\boldsymbol{U}_i - \bar{\boldsymbol{U}}_{r-1}) &= \{\underline{\boldsymbol{A}}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})\}^{\top}(\underline{\boldsymbol{A}}\underline{\boldsymbol{S}}_{r-1}\underline{\boldsymbol{A}}^{\top})^{-1}\underline{\boldsymbol{A}}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}) \\ &= (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^{\top}(\underline{\boldsymbol{S}}_{r-1})^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1}). \end{aligned} \tag{A19}$$

This shows that the D-optimal selection function remains unchanged under affine transformations and hence, FSM with the D-optimal selection function is affine invariant.

(b) The in-sample symmetry of the data essentially implies that if $\boldsymbol{X}$ belongs to the sample, then $-\boldsymbol{X}$ also belongs to the sample. Moreover, by the assumption of a continuous data generating distribution, with probability 1, the covariate values are different up to reflection. Now, consider an even transformation $g(\cdot)$, i.e., $g(-\boldsymbol{X}) = g(\boldsymbol{X})$. With two groups of equal sizes, the SOM consists of successive random permutations of the 'chunk' $(1, 2)$. By Theorem 4.1, for the first pair of stages of selection, the objective function (to maximize) is given by

$$(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}})^{\top}(\underline{\boldsymbol{S}}_{\text{full}})^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{\text{full}}) = \boldsymbol{X}_i^{\top}(\underline{\boldsymbol{S}}_{\text{full}})^{-1}\boldsymbol{X}_i. \tag{A20}$$

It follows that, if a unit in the sample with covariate $\boldsymbol{X}^{(1)}$ maximizes the objective function in Equation A20, then so does the unit with covariate $-\boldsymbol{X}^{(1)}$. Moreover, due to the continuous data generating distribution, with probability 1, these are the only two units that

maximize this objective function. Therefore, if treatment 1 selects the unit with covariate $\boldsymbol{X}^{(1)}$, treatment 2 selects the unit with covariate $-\boldsymbol{X}^{(1)}$, and vice-versa. This preserves exact balance on $g(\boldsymbol{X})$.

Now, consider the next pair of stages. Without loss of generality, suppose treatment 1 had chosen a unit with covariate $\boldsymbol{X}^{(1)}$ and treatment 2 had chosen a unit with covariate $-\boldsymbol{X}^{(1)}$ in their respective previous choices. Also, without loss of generality, assume that in this pair of stages, treatment 1 gets to choose first. By Theorem 4.1, treatment 1 aims to maximize,

$$
\begin{aligned}
&(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1})^\top (\underline{\boldsymbol{S}}^*_{r-1})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}^*_{r-1}) \\
&= \{\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}^\top \Big\{\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}^\top_{\text{full}}\underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top}\Big\}^{-1} \{\boldsymbol{X}_i - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}.
\end{aligned}
\tag{A21}
$$

Also, during treatment 2's turn in this pair of stages, it tries to maximize

$$
\begin{aligned}
&(\boldsymbol{X}_i + \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)})^\top \Big\{\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}^\top_{\text{full}}\underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top}\Big\}^{-1} (\boldsymbol{X}_i + \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}) \\
&= \{(-\boldsymbol{X}_i) - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}^\top \Big\{\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top} + \frac{\epsilon}{N}\underline{\boldsymbol{X}}^\top_{\text{full}}\underline{\boldsymbol{X}}_{\text{full}} - (1+\epsilon)\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)\top}\Big\}^{-1} \{(-\boldsymbol{X}_i) - \frac{1}{1+\epsilon}\boldsymbol{X}^{(1)}\}.
\end{aligned}
\tag{A22}
$$

Equations A21 and A22 imply that if treatment 1 chooses a unit with covariate value $\boldsymbol{X}^{(2)}$, then with probability 1, treatment 2 chooses the unit with covariate value $-\boldsymbol{X}^{(2)}$, and vice versa. This shows that, at the end of the second pair of stages in the SOM, exact mean balance on $g(\boldsymbol{X})$ is preserved. Proceeding in this manner it follows that, at the end of the selection process, with probability 1, both the treatment groups will have exact balance on $g(\boldsymbol{X})$. This completes the proof.

□

It follows from Theorem A2(a) that, for any SOM, the choices made by each treatment group remain unchanged even if the covariate vectors are transformed via an affine transformation

(e.g., changing the units of measurement of the covariates). Therefore, the FSM with the D-optimal selection function self-standardizes the covariates. In addition, if the covariate vector is symmetrically distributed in the sample, then by Theorem A2(b), the FSM exactly balances even transformations such as the second, fourth order moments, and the pairwise products of the centerd covariates. An implication of Theorem A2(b) is that, for covariates drawn from symmetric continuous distributions (such as the Normal, t, and Laplace distributions), the FSM tends to balance all these transformations due to the approximate symmetry of the covariates in the sample. The choice of the D-optimal selection function is thus robust in the sense that it allows the FSM to balance a family of transformations of the covariate vector by design, without explicitly including them in the assumed linear model nor requiring the specification of tuning parameters.

C.2 Connection to A-optimality

The original FSM used a criterion based on A-optimality as the selection function (see Morris 1979). In this section, we compare the A-and D-optimal selection functions. The A-optimal selection function requires prespecifying a *policy matrix* $\underline{\boldsymbol{P}}_{p\times(k+1)}$ and a corresponding vector of *policy weights* $\boldsymbol{w}_{p\times 1}$. Here, $\underline{\boldsymbol{P}}$ transforms the original vector of regression coefficients to a vector of p linear combinations that are of policy interest, and $\boldsymbol{w}$ assigns weights to each combination according to their importance. Thus, the A-optimal selection function requires $p(k+2)$ tuning parameters.

If treatment 1 gets to choose at the rth stage, then this criterion selects the unit that minimizes the resulting trace$\Big\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r,i}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r,i})^{-1}\Big\}$, where $\underline{\boldsymbol{T}} = \underline{\boldsymbol{P}}^{\top}\text{diag}(\boldsymbol{w})\underline{\boldsymbol{P}}$. Proposition A3 shows an equivalent characterization of the A-optimal selection function.

Proposition A3. Let treatment 1 be the choosing group at the rth stage. Assume that $\tilde{n}_{r-1} \geq 1$ and $\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1}$ is invertible. The A-optimal selection function chooses unit i' with covariate vector $\boldsymbol{X}_{i'} \in \mathbb{R}^k$, where $i' \in \arg\max_{i\in\mathcal{R}_{r-1}} \frac{(1,\boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}{1+(1,\boldsymbol{X}_i^{\top})(\underline{\tilde{\boldsymbol{X}}}_{r-1}^{\top}\underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\left(\begin{smallmatrix}1\\ \boldsymbol{X}_i\end{smallmatrix}\right)}$.

The A-optimality criterion provides a family of selection functions depending on $\underline{\boldsymbol{P}}$ and $\boldsymbol{w}$.

For some choices of $\underline{\boldsymbol{P}}$ and $\boldsymbol{w}$, the selection function is not affine invariant , e.g., $\underline{\boldsymbol{P}} = \boldsymbol{I}$ and $\boldsymbol{w} = (1, 1, ..., 1)^\top$, while for other choices it is, e.g., $\underline{\boldsymbol{P}} = \underline{\tilde{\boldsymbol{X}}}_{\text{full}}$ and $\boldsymbol{w} = (1, 1, ..., 1)^\top$. In particular, the A-optimal selection function with $\underline{\boldsymbol{P}} = \underline{\tilde{\boldsymbol{X}}}_{\text{full}}$ and $\boldsymbol{w} = (1, 1, ..., 1)^\top$ is closely related to the D-optimal selection function. To see this, consider a case where in the selection process, the design matrices in each treatment group scale similarly relative to the design matrix in the full sample, i.e., $(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} = c_r(\underline{\tilde{\boldsymbol{X}}}_{\text{full}}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{full}})^{-1}$ for some constant $c_r > 0$. In this case, the A-optimal selection function chooses unit i' such that $i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix} \iff i' \in \arg\max_{i \in \mathcal{R}_{r-1}} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})^\top (\underline{\boldsymbol{S}}_{r-1})^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{r-1})$, which is equivalent to the D-optimal selection function. Hence, in this case, the FSM under the D-optimal and A-optimal selection functions make similar choices of units.

Proof of Proposition A3

Proof. The A-optimal selection function aims to minimize

$$\begin{aligned}
&\text{trace}\left\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r,i}^\top \underline{\tilde{\boldsymbol{X}}}_{r,i})^{-1}\right\} && \text{(A23)}\\
&= \text{trace}\left[\underline{\boldsymbol{T}}\{\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1} + \begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}(1, \boldsymbol{X}_i^\top)\}^{-1}\right] \\
&= \text{trace}\left\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1} - \underline{\boldsymbol{T}}\frac{(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}}{1 + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}\right\} \\
&= \text{trace}\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\} - \text{trace}\left\{\underline{\boldsymbol{T}}\frac{(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}}{1 + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}\right\} \\
&= \text{trace}\{\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\} - \text{trace}\left\{\frac{(1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}{1 + (1, \boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}\right\} && \text{(A24)}
\end{aligned}$$

□

Here the second equality holds due to the Sherman-Morrison-Woodbury formula, the third and fourth equality hold due to the linearity and cyclicality of trace$(\cdot)$, respectively. Equation A24 shows that it is equivalent to maximize trace$\left\{\frac{(1,\boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\underline{\boldsymbol{T}}(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}{1+(1,\boldsymbol{X}_i^\top)(\underline{\tilde{\boldsymbol{X}}}_{r-1}^\top \underline{\tilde{\boldsymbol{X}}}_{r-1})^{-1}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}}\right\}$. This completes the proof.

D Optimal covariance design theorem and D-optimality

In this section, we focus on the setting with $G = 2$ treatment groups. Under a model-based approach, we first connect the notion of covariate balance to efficiency using the optimal covariance design theorem (Morris and Hill 2000, see also Chattopadhyay et al. 2021)

Theorem A4. Consider the linear regression model $Y_i^{\text{obs}} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i + \tau \mathbb{1}(Z_i = 2) + \epsilon_i$, where ϵ_is are the uncorrelated error terms with mean zero and variance σ^2. Let $\hat{\tau}_{\text{OLS}}$ be the ordinary least squares estimator of τ. Then,

$$\text{Var}(\hat{\tau}_{\text{OLS}}) = \frac{\sigma^2}{N s_2^2 (1 - R^2)},$$

where $s_2^2 = \frac{n_1 n_2}{N^2}$ and R^2 is the square of the multiple correlation coefficient of $\mathbb{1}(Z_i = 2)$ with the covariates.

Here, $\hat{\tau}_{\text{OLS}}$ is used to estimate the average treatment effect of treatment 2, relative to treatment 1. Theorem A4 implies that, under this model, the most efficient design minimizes R^2. In other words, the optimal design satisfies $R^2 = 0$ (if feasible), which equivalently means that the covariates $\boldsymbol{X}_i$ are exactly mean-balanced across the two treatment groups. Indeed, the optimality of this design is optimal depends heavily on the correctness of the outcome model. With model misspecification, this design may no longer be efficient. For instance, if the outcome model is linear in second-order transformations of the covariates, the design may perform poorly due to potential lack of balance on these transformations. In this sense, deterministic optimal designs lack robustness against model misspecification.

Next, we consider the global D-optimal design, i.e., the design that selects the D-optimal assignment among all possible assignments. If there are multiple D-optimal assignments, one of them is chosen randomly by the design. Proposition A5 shows that, with $k = 1$ covariate, the global D-optimal design aims to balance the means of the covariate exactly between the two treatment groups.

Proposition A5. Consider the linear model $Y_i^{\text{obs}} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i + \tau \mathbb{1}(Z_i = 2) + \epsilon_i$, where ϵ_is are the uncorrelated error terms with mean zero and variance σ^2. Under this model, the global D-optimal design minimizes $|\bar{X}_1 - \bar{X}_2|$, where $\bar{X}_1$ and $\bar{X}_2$ are the means of X_i in treatment groups 1 and 2, respectively.

Proposition A5 and Theorem A4 imply that, if the outcome model is linear in the covariates and the treatment, then the global D-optimal design is the most efficient.

Proof of Proposition A5

By definition, the D-optimal design maximizes $\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}})$, where $\underline{\boldsymbol{D}} = (\mathbf{1}, \boldsymbol{X}, \boldsymbol{Z})$ is the design matrix. Without loss of generality, we assume that the covariates are scaled so that their variance in the full sample is 1, i.e., $\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_{\text{full}})^2 = 1$. Then,

$$\begin{aligned}
\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}}) &= \det \begin{pmatrix} N & N\bar{X}_{\text{full}} & n_1 \\ N\bar{X}_{\text{full}} & N + N\bar{X}_{\text{full}} & n_1\bar{X}_1 \\ n_1 & n_1\bar{X}_1 & n_1 \end{pmatrix} \\
&= N^2 n_1 - N n_1^2 - N n_1^2(\bar{X}_{\text{full}}^2 + \bar{X}_1^2 - 2\bar{X}_{\text{full}}\bar{X}_1) \\
&= N n_1 n_2 - \frac{n_1^2 n_2^2}{N}(\bar{X}_1 - \bar{X}_2)^2, \qquad \text{(A25)}
\end{aligned}$$

where the last equality holds since $\bar{X}_{\text{full}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{N}$. Thus, maximizing $\det(\underline{\boldsymbol{D}}^\top \underline{\boldsymbol{D}})$ is equivalent to minimizing $|\bar{X}_1 - \bar{X}_2|$. This completes the proof.

E Algorithms for constructing an SOM

E.1 The SCOMARS algorithm

Consider a setting with $G = 2$ treatment groups of arbitrary sizes n_1 and n_2. Let W_r be the binary indicator for selection of group 1 stage r, $r \in \{1, 2, \ldots, N\}$, with $p_r := P(W_r = 1)$ being the marginal probability of selection at stage r. Write $S_r := \sum_{j=1}^{r} W_j$ and $F_r := \mathbb{E}(S_r) = \sum_{j=1}^{r} p_j$. A treatment assignment is sequentially controlled if $|S_r - F_r| < 1$ for all

$r \in \{1, 2, \ldots, N\}$.

The SCOMARS algorithm proceeds as follows:

- Stage 1, $P(W_1 = 1) = p_1$.
- Stage $r \geq 2$, $P(W_r = 1 | S_{r-1} = s_{r-1}) = P\Big\{U \leq \frac{p_r - \max(0, s_{r-1} - F_{r-1})}{1 - |s_{r-1} - F_{r-1}|}\Big\}$, where $U \sim$ Unif$(0, 1)$.

This algorithm satisfies the sequentially controlled condition, $|S_r - F_r| < 1$ for all $r \in \{1, 2, \ldots, N\}$ (Morris 1983). It is Markovian because the probability of selection at stage r depends solely on stage $r - 1$.

E.2 SOM for multi-group experiments

We first define the randomized chunk algorithm for generating an SOM for multi-group experiments with equal group sizes.

Definition 1 (Randomized chunk algorithm)**.** Suppose $n_1 = n_2 = ... = n_G$. The randomized chunk algorithm generates an SOM by generating and stacking $\frac{N}{G}$ independent random permutations of the 'chunk' $(1, 2, ..., G)$.

For example, with $N = 12$, $g = 3$, $n_1 = n_2 = n_3 = 4$, one instance of an SOM generated using randomized chunk is $(\underbrace{2, 1, 3}, \underbrace{1, 2, 3}, \underbrace{2, 1, 3}, \underbrace{2, 3, 1})^\top$.

The following proposition shows that the randomized chunk algorithm is sequentially controlled.

Proposition A6. For $G \geq 2$ and $n_1 = n_2 = ... = n_G$, the randomized chunk algorithm satisfies $|S_{ig} - F_{ig}| \leq \frac{G-1}{G} < 1$ for all $g \in \{1, 2, ..., G\}$.

Proof. Let S_{ig} and F_{ig} be the same as defined in Section 8.1 ($i \in \{1, 2, ..., N\}$, $g \in \{1, 2, ..., G\}$). For equal sized treatment groups, $F_{ig} = \frac{i}{G}$. Now, without loss of generality, it suffices to show that $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., N\}$. Consider the first chunk in the SOM, which is a random permutation of $(1, 2, ..., G)$. If treatment 1 appears in position $i^* \in \{1, 2, ..., G\}$

the permutation ($j \in \{1, 2, ..., G\}$), then

$$|S_{i1} - F_{i1}| = \begin{cases} \frac{i}{G} & \text{if } i \in \{1, ..., i^* - 1\} \\ 1 - \frac{i}{G} & \text{if } i \in \{i^*, ..., G\}. \end{cases} \tag{A26}$$

In each case, $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., G\}$. Moreover, since $|S_{G1} - F_{G1}| = 0$, the SOM restarts itself after the first chunk. Hence, we can conclude that $|S_{i1} - F_{i1}| \leq \frac{G-1}{G}$ for all $i \in \{1, 2, ..., G\}$. This completes the proof.

□

Below we describe two algorithms to generate an SOM for multi-group experiments and show that they are sequentially controlled. The key idea in these algorithms is the formation of 'supergroups', i.e., combination of one or more treatment groups. For example, with $g = 3$, $n_1 = 10, n_2 = 20, n_3 = 30$, one can consider two supergroups, namely $\{1, 2\}$ of size $10 + 20 = 30$ and $\{3\}$ of size 30.

Theorem A7. For $1 \leq G_1 \leq G - 1$, let $n_1 = n_2 = ... = n_{G_1} \neq n^{(1)}$, and $n_{G_1+1} = n_{G_1+2} = ... = n_G = n^{(2)}$, where $n^{(1)} \neq n^{(2)}$. Consider the following three-stage algorithm.

1. Run SCOMARS with supergroups $\{1, ..., G_1\}$ and $\{G_1 + 1, ..., G\}$ to generate an SOM at the supergroup level.

2. Consider the locations of the SOM in step 1 where supergroup $\{1, ..., G_1\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

3. Consider the locations of the SOM in step 1 where supergroup $\{G_1 + 1, ..., G\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

The above SOM generating algorithm is sequentially controlled.

We first prove a special case of Theorem A7, given below in Lemma A2

Lemma A2. The algorithm in Theorem A7 is sequntially controlled for the special case of $G_1 = 1$.

Proof. The first step of the algorithm in Theorem A7 runs SCOMARS with treatment group 1 and the supergroup $\{2, 3, ..., G\}$. Thus, the first step itself determines the locations of the SOM where treatment 1 gets to choose. Since SCOMARS is sequentially controlled, we immediately have $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$.

It remains to show that for $g \in \{2, 3, ..., G\}$, $|S_{ig} - F_{ig}| < 1$ for all $i \in \{1, 2, ..., N\}$. By symmetry, it suffices to show this for $g = 2$. Now, the randomized chunk algorithm on the supergroup $\{2, 3, ..., G\}$ determines the locations of the SOM where treatment 2 gets to choose. We will prove the result by first mapping this SOM to an SOM where treatment 1 is absent, and then by using the sequential controlled property of randomized chunk.

Let us first denote $1 \leq r_1 < r_2 < ... < r_{n_1-1} < r_{n_1} \leq N$ as the stages or locations of the SOM where treatment 1 gets to choose. We consider the following cases,

(i) Case-1: $i \in \{1, 2, ..., r_1 - 1\}$. In this case, by stage i, treatment 1 has not made any choices. Now,

$$\begin{aligned} |S_{i2} - F_{i2}| &= |S_{i2} - \frac{in^{(2)}}{N}| \\ &\leq |S_{i2} - \frac{i}{G-1}| + |\frac{i}{G-1} - \frac{in^{(2)}}{N}| \\ &\leq \frac{G-2}{G-1} + i\frac{n_1}{N(G-1)} \\ &< \frac{G-2}{G-1} + \frac{1}{G-1} = 1. \end{aligned} \tag{A27}$$

Here the first inequality holds due to triangle inequality. To see that second inequality, consider a new experiment with treatment groups $\{2, ..., G\}$ of size $n^{(2)}$ each and an SOM generated by randomized chunk as in the second step of the algorithm in Theorem A7. Let

$\tilde{S}_{i2}$ be the number of selections made by treatment 2 up to stage i in this new experiment and $\tilde{F}_{i2} = \frac{i}{G-1}$ be its expectation. By Proposition A6, $|\tilde{S}_{i2} - \tilde{F}_{i2}| \leq \frac{G-2}{G-1}$. Now, $|S_{i1} - \frac{i}{G-1}| = |\tilde{S}_{i1} - \frac{i}{G-1}|$, which gives us the second inequality. Finally, the last inequality holds since $\frac{in_1}{N} = F_{i1} < 1$.

(ii) Case-2: $i \in \{r_t, r_t + 1, ..., r_{t+1} - 1\}$ for some $t \in \{1, 2, ..., n_1 - 1\}$. In this case, by stage i, treatment 1 has made exactly t choices. Now,

$$\begin{aligned} |S_{i2} - F_{i2}| &= |S_{i2} - \frac{in^{(2)}}{N}| \\ &\leq |S_{i2} - \frac{i-t}{G-1}| + |\frac{i-t}{G-1} - \frac{in^{(2)}}{N}| \\ &\leq \frac{G-2}{G-1} + \frac{1}{G-1}|t - \frac{in_1}{N}| \\ &< \frac{G-2}{G-1} + \frac{1}{G-1} = 1. \end{aligned} \tag{A28}$$

Here, the first inequality is due to triangle inequality. To see the second inequality, we again consider the new experiment described in Case-1. Notice that, $|S_{i2} - \frac{i-t}{G-1}| = |\tilde{S}_{(i-t)2} - \tilde{F}_{(i-t)2}| \leq \frac{G-2}{G-1}$, where the last inequality holds by Proposition A6. Finally, the final inequality in Equation A28 holds since $|t - \frac{in_1}{N}| = |S_{i1} - F_{i1}| < 1$. This completes the proof of the lemma. □

We now prove Theorem A7.

Proof. We first show that, for $g \in \{1, 2, ..., G_1\}$,

$$|S_{ig} - F_{ig}| < 1 \quad \forall i \in \{1, 2, ..., N\}. \tag{A29}$$

To show this, we consider steps 1 and 2 of the algorithm as these two steps are sufficient to determine the location of treatments $1, ..., G_1$ in the SOM. We note that, steps 1 and 2 generate an SOM for an experiment with $G_1 + 1$ treatment groups, namely supergroup $\{G_1 + 1, ..., G\}$ (of size $(G - G_1)n^{(2)}$) and groups $1, 2, ..., G_1$ (each of size $n^{(1)}$). Thus, by

Lemma A2, it follows that Equation A29 holds for $g \in \{1, 2, ..., G_1\}$.

To show that Equation A29 holds for $g \in \{G_1 + 1, ..., G\}$, we first notice that steps 2 and 3 of the algorithm are completely independent and hence can be performed in any order. Therefore, by changing the order of steps 2 and 3 and applying the same argument as before, we get that Equation A29 holds for $g \in \{G_1 + 1, ..., G\}$. This completes the proof of the theorem. □

Theorem A8. Let $G_1, ..., G_m$ be such that $1 \leq G_j \leq G - 1$ for all $j \in \{1, 2, ..., m\}$ and $G_1 + G_2 + ... + G_m = G$. Moreover, for $j \in \{1, 2, ..., m\}$, let $n^{(j)}$ be the group size of G_j many treatment groups, with $n^{(1)}G_1 = n^{(2)}G_2 = ... = n^{(m)}G_m$. Denote the collection of G_j treatment groups with group sizes $n^{(j)}$ as supergroup $\mathcal{G}_j$. Consider the following multi-stage algorithm.

1. Run randomized chunk on supergroups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m$ to generate an SOM at the supergroup level.

2. For $j \in \{1, 2, ..., m\}$, consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_j$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

The above SOM generating algorithm is sequentially controlled.

To prove this theorem, we first use the following Lemma.

Lemma A3. Let $n_1 = n_2 = ... = n_G = n$. Consider the following SOM generating algorithm.

1. Consider the supergroups $\{1\}$ (of size n) and $\{2, 3, ..., G\}$ (of size $(G-1)n$). Generate an SOM at the superpopulation level using SCOMARS.

2. Consider the locations of the SOM in step 1 where supergroup $\{2, 3, ..., G\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

This algorithm is equivalent to the randomized chunk algorithm.

Below we prove this lemma.

Proof. To show that the algorithm is equivalent to randomized chunk, we have to show that it generates a random permutation of $(1, 2, ..., G)$ for the first G stages, a fresh random permutation of $(1, 2, ..., G)$ for the next G stages, and so on. Since the locations of groups $\{2, ..., G\}$ are chosen using randomized chunk, it thus suffices to show that, treatment 1 gets to choose once (in a random location) in the first G stages, once in the next G stages, and so on.

We use the notation as in Section E.1. Now, suppose among the first G stages, treatment 1 gets to choose at stage r^* first. Notice that r^* cannot be greater than G as

$$P(W_G = 1|S_{G-1} = 0) = P\Big\{U \leq \frac{\frac{1}{G} - \max(0, 0 - F_{G-1})}{1 - |0 - F_{G-1}|}\Big\} = P\Big\{U \leq \frac{1}{G - (G-1)}\Big\} = 1. \tag{A30}$$

Now, for $r \in \{1, 2, ..., r^* - 1\}$ we have,

$$P(W_r = 1|S_{r-1} = 0) = P\Big\{U \leq \frac{\frac{1}{G} - \max(0, 0 - F_{r-1})}{1 - |0 - F_{r-1}|}\Big\} = P\Big\{U \leq \frac{1}{G - (r-1)}\Big\} = \frac{1}{G - (r-1)}. \tag{A31}$$

For $r^* + 1 \leq r \leq G$,

$$P(W_r = 1|S_{r-1} = 1) = P\Big\{U \leq \frac{p_r - \max(0, 1 - F_{r-1})}{1 - |1 - F_{r-1}|}\Big\} = P\Big\{U \leq \frac{\frac{1}{G} - 1 + \frac{r-1}{G}}{\frac{r-1}{G}}\Big\} = 0. \tag{A32}$$

Finally,

$$P(W_{G+1} = 1|S_G = 1) = P\Big\{U \leq \frac{p_{G+1} - \max(0, 1 - F_G)}{1 - |1 - F_G|}\Big\} = P\Big(U \leq \frac{1}{G}\Big) = \frac{1}{G}. \tag{A33}$$

Therefore, by Equation A32, if treatment 1 selects at the r^*th stage, it never selects again $2, 3, ..., G$. Also, by Equation A31, before the r^*th stage, the conditional probabilities of treatment 1 selecting are same as what it would have been under random permutation of the group labels. Finally, by Equation A33, the process restarts itself at the $(G+1)$th stage, which is equivalent to starting a fresh new random permutation of the group labels. This completes the proof of the lemma. □

We now prove Theorem A8.

Proof. By the symmetry of the problem, it suffices to show that $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$. Without loss of generality, we assume that $\mathcal{G}_1 = \{1, 2, ..., G_1\}$, which implies that treatment 1 belongs to supergroup $\mathcal{G}_1$. Now, it suffices to focus on the following to steps of the algorithm:

1. Run randomized chunk on supergroups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_m$ to generate an SOM at the supergroup level.

2. Consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_1$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

This is because, these two steps completely determine the locations of treatment 1 in the SOM. By Lemma A3, these two steps can be equivalently performed as follows.

1. Consider the supergroups $\mathcal{G}_1$ (of size $n^{(1)}G_1$) and $\{\mathcal{G}_2, ..., \mathcal{G}_m\}$ (of size $(m-1)n^{(1)}G_1$). Generate an SOM at this supergroup level using SCOMARS.

2. Consider the locations of the SOM in step 1 where supergroup $\{\mathcal{G}_2, ..., \mathcal{G}_m\}$ chooses. Then, use randomized chunk to obtain the selection orders at the levels of $\mathcal{G}_j$ in those locations.

3. Consider the locations of the SOM in step 1 where supergroup $\mathcal{G}_1$ chooses. Then, use

randomized chunk to obtain the selection orders at the levels of the original groups in those locations.

We note that this above algorithm is exactly equivalent to the SOM generating algorithm in Theorem A7 for an experiment with G_1+m-1 treatment groups, namely, $1, 2, ..., G_1, \mathcal{G}_2, \mathcal{G}_3, ..., \mathcal{G}_m$. Thus, by Theorem A7, we have $|S_{i1} - F_{i1}| < 1$ for all $i \in \{1, 2, ..., N\}$. □

F FSM for stratified experiments

In this section, we discuss two potential approaches to use an FSM for stratified experiments. We consider stratified experiments where the treatment group sizes within each stratum are set by the investigator beforehand. To accommodate the FSM to such experiments, we again need to carefully construct an SOM. In particular, we append the SOM with an additional column of stratum labels, indicating which stratum the treatment group selects from at each stage of the selection process. This column of stratum labels is specified in such a way that the resulting SOM satisfies the group size requirements within each stratum.

Conceptually, the most straightforward approach is to generate a separate SOM for each stratum. This is equivalent to setting the column of stratum labels as $(\underbrace{1, ..., 1}_{m_1}, \underbrace{2, ..., 2}_{m_2}, ..., \underbrace{S, ..., S}_{m_S})^\top$, where S is the number of strata and m_s is the size of sth stratum, $s \in \{1, 2, ..., S\}$. This approach is easy to implement and can be useful if, e.g., data on each stratum is available at different stages of the experiment, akin to a sequential experiment. However, in this approach, the treatment groups only get to explore the covariate space of a single stratum for a number of successive stages of selection and hence may not make the most efficient choices. We address this issue with an alternative approach. For ease of exposition, we consider two strata: 1 and 2. Let n_{1g} and n_{2g} be the (fixed) sizes of treatment group $g \in \{1, 2, ..., G\}$ in strata 1 and 2, respectively, where $n_{1g} + n_{2g} = n_g$. In this approach, we first generate a usual SOM with group sizes $n_1, ..., n_G$. For $g \in \{1, 2, ..., G\}$, we then select the order of the strata that treatment g chooses from by running a SCOMARS algorithm with group

sizes n_{1g} and n_{2g}. By allowing the treatment groups to select units from different strata in a balanced manner, this approach mimics the unstratified FSM where the covariate space of the entire sample is explored for choosing units. Also, by design, this approach satisfies the size requirement of each treatment group within each stratum.

G FSM for sequential experiments

In this section, we describe our approach to using the FSM for sequential experiments. Suppose treatment 1 gets to choose at the first stage of selection for the new batch. Let $\underline{\tilde{\boldsymbol{X}}}_{\text{old}}$ be the design matrix based on units already assigned to treatment 1. Also, for each unit i in the new batch, let $\underline{\tilde{\boldsymbol{X}}}_{\text{new},i} := \begin{pmatrix} \underline{\tilde{\boldsymbol{X}}}_{\text{old}} \\ (1, \boldsymbol{X}_i^\top) \end{pmatrix}$ be the resulting design matrix in treatment group 1 if unit i is selected. Treatment 1 selects the unit that maximizes $\det(\underline{\tilde{\boldsymbol{X}}}_{\text{new},i}^\top \underline{\tilde{\boldsymbol{X}}}_{\text{new},i})$. In other words, we use the design matrix based on all the units already assigned to the choosing treatment group to evaluate the D-optimal selection function for each unit in the new batch, and select the unit that maximizes the selection function. By carrying over the existing design matrix to the new batch, this approach tends to correct for any existing covariate imbalances.

H A simulation study

H.1 Setup

We now compare the performance of the FSM to complete randomization and rerandomization in a simulation study. Here, $N = 120$, $G = 2$, $n_1 = n_2 = 60$, and $k = 6$. The covariates are generated following the design of Hainmueller (2012):

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix} \right\}, \; X_4 \sim \text{Unif}(-3, 3), \; X_5 \sim \chi^2_1, \; X_6 \sim \text{Bernoulli}(0.5). \tag{A34}$$

In this design, X_4, X_5, and X_6 are mutually independent and separately independent of $(X_1, X_2, X_3)^\top$. We draw a sample of 120 units once from the data generating mechanism in (A34). Conditional on this sample, we compare four different assignment methods, namely

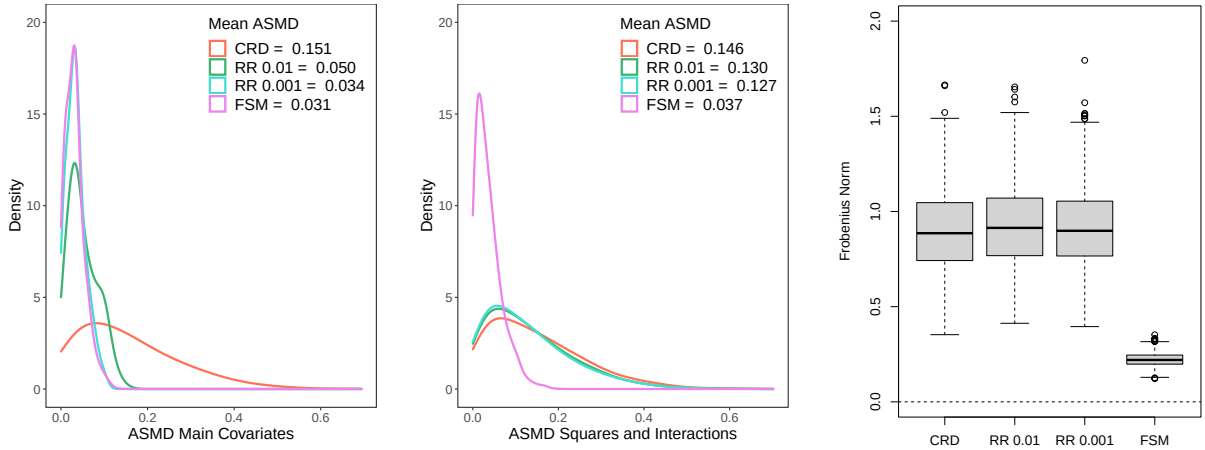
a completely randomized design (CRD), rerandomization with 0.01 acceptance rate (RR 0.01), rerandomization with 0.001 acceptance rate (RR 0.001), and the FSM. Both RR 0.01 and RR 0.001 use as rerandomization criteria the Mahalanobis distance between the two treatment groups on the original covariates. The FSM uses a linear potential outcome model on the original covariates and the D-optimal selection function. For each design we draw 800 independent assignments. The assignments under the FSM are generated using the open source R package `FSM` available on CRAN. The total runtime of the FSM for the 800 simulated experiments was about one and a half minutes on a Windows 64-bit computer with an Intel(R) Core i7 processor. See Chattopadhyay et al. (2021) for detailed step-by-step instructions and vignettes on the use of FSM package.

H.2 Balance

We evaluate balance on the main and transformed covariates. Figures A1(a) and A1(b) show density plots of the Absolute Standardized Mean Differences (ASMD; Rosenbaum and Rubin 1985, Stuart 2010) of the six main covariates and their second-order transformations (including squares and pairwise products), respectively. A smaller ASMD for a covariate indicates better mean-balance on that covariate between the two treatment groups. Figure A1(a) indicates that both rerandomization methods improve balance on the means of the original covariates over CRD. As expected, the ASMD distribution under RR 0.001 is more concentrated than that of RR 0.01, with 32% smaller mean ASMD than RR 0.01. Both the FSM and RR 0.001 have similar distributions of the ASMD with FSM having moderately (9%) smaller mean ASMD. See Table A10 for a comparison of the average ASMD of each covariate.

Figure A1(b) shows that the imbalances of covariate transformations are substantially smaller with the FSM than with CRD, RR 0.01, and RR 0.001. In fact, the FSM achieves a 70% reduction in the mean ASMD with respect to RR 0.001. Thus, although the FSM and RR 0.001 exhibit comparable balance in terms of the main covariates, the FSM balances

Figure A1: Panels (a) and (b) show distributions of absolute standardized mean differences (ASMD) of the main covariates and all their second-order transformations across 800 randomizations. For each plot, the legend presents the average ASMD across simulations for the four methods. Panel (c) shows distributions of discrepancies between the correlation matrices of the covariates in the treatment and the control group (as measured by the Frobenius norm, $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$). On average the FSM achieves better covariate balance. In terms of the main covariates, the FSM marginally outperforms RR 0.001. In terms of the second-order transformations and correlation matrices, the FSM substantially outperforms RR 0.001.



(a) Main covariates (b) Squares and pairwise products (c) Frobenius norm

these transformations of the covariates much better than RR 0.001. This highlights the improved robustness of the FSM against model misspecification. Moreover, reducing the tuning parameter of rerandomization from 0.01 to 0.001 yields only 2% improvement in the mean ASMD.[1] In Figure A1(b), both RR 0.01 and RR 0.001 often produce ASMD larger than 0.1, and in some cases, larger than 0.5, indicative of substantial imbalances on these covariate transformations.

For each method, we also compare balance in the overall correlation structure of the covariates. Figure A1(c) shows the boxplots of the distributions of $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$. The FSM outperforms the other three designs with at least 75% smaller average $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$. In particular, among the 800 randomizations, the highest value of $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$ under FSM is smaller than the corresponding lowest value under the other three designs, indicating that in terms

[1] In fact, for some covariate transformations, reducing this tuning parameter exacerbates imbalance (see Table A11).

of the correlation structure (and hence the interactions) of the covariates, the least balanced realization of the 800 FSMs exhibits better balance than the best balanced realization of the 800 complete randomizations and rerandomizations.

H.3 Efficiency

We now compare the efficiency of the methods under both model- and randomization-based approaches to inference. Under the model-based approach, we consider a potential outcome model where $\mathbb{E}\{Y_i(g)|\boldsymbol{X}_i\}$ is linear in $\boldsymbol{X}_i$ (Model A1), and another model where $\mathbb{E}\{Y_i(g)|\boldsymbol{X}_i\}$ is linear in $\boldsymbol{X}_i$ and all its second-order transformations (Model A2). In each case, we assume homoscedasticity, i.e., $\text{Var}\{Y_i(g)|\boldsymbol{X}_i\} = \sigma^2$. For each potential outcome model, we fit the corresponding observed outcome model by OLS and estimate $\text{PATE}_{2,1}$ using the regression imputation method described in Section 5.

More concretely, consider a specific treatment assignment vector $\boldsymbol{Z}$. Under Model A1, we fit a linear regression model $Y_i^{\text{obs}} = (1, \boldsymbol{X}_i)^\top \boldsymbol{\beta}_g + \epsilon_{ig}$ in treatment group $g \in \{1, 2\}$ and estimate $\text{PATE}_{2,1}$ by the regression imputation estimator $\widehat{\text{PATE}}_{2,1} = \hat{\boldsymbol{\beta}}_2^\top \bar{\tilde{\boldsymbol{X}}} - \hat{\boldsymbol{\beta}}_1^\top \bar{\tilde{\boldsymbol{X}}}$, where $\bar{\tilde{\boldsymbol{X}}}^\top = \frac{1}{N}\sum_{i=1}^N (1, \boldsymbol{X}_i^\top)$. The model-based standard error of this estimator is $\text{SE}_{\boldsymbol{Z}} = \sigma\sqrt{\bar{\tilde{\boldsymbol{X}}}^\top \{(\underline{\tilde{\boldsymbol{X}}}_{1,\boldsymbol{Z}}^\top \underline{\tilde{\boldsymbol{X}}}_{1,\boldsymbol{Z}})^{-1} + (\underline{\tilde{\boldsymbol{X}}}_{2,\boldsymbol{Z}}^\top \underline{\tilde{\boldsymbol{X}}}_{2,\boldsymbol{Z}})^{-1}\} \bar{\tilde{\boldsymbol{X}}}}$, where $\underline{\tilde{\boldsymbol{X}}}_{g,\boldsymbol{Z}}$ is the design matrix in group g, for the given treatment assignment $\boldsymbol{Z}$.

Now, for a design d, the average and maximum model-based standard error relative to the FSM is given by $\frac{\frac{1}{M}\sum_{r=1}^M \text{SE}_{\boldsymbol{Z}_d^{(r)}}}{\frac{1}{M}\sum_{r=1}^M \text{SE}_{\boldsymbol{Z}_{\text{FSM}}^{(r)}}}$ and $\frac{\max_r \text{SE}_{\boldsymbol{Z}_d^{(r)}}}{\max_r \text{SE}_{\boldsymbol{Z}_{\text{FSM}}^{(r)}}}$, respectively, where $\boldsymbol{Z}_d^{(1)}, ..., \boldsymbol{Z}_d^{(M)}$ are M independent assignment vectors generated under design d, and $\boldsymbol{Z}_{\text{FSM}}^{(1)}, ..., \boldsymbol{Z}_{\text{FSM}}^{(M)}$ are generated under the FSM. These measures do not depend on σ^2 and can be computed exactly from the observed data. Tables A4(a) and A4(b) show the average and maximum model-based standard error (SE) of the regression imputation estimator relative to the FSM across $M =$ 800 randomizations under the two models.

Table A4: Average and maximum model-based standard errors relative to the FSM across randomizations. Under Model A1 (linear model on the main covariates), the FSM and RR exhibit similar performance, improving over CRD. Under Model A2 (linear model on the main covariates and their second-order transformations), the FSM is considerably more efficient than both CRD and RR.

(a) Model A1

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
Average SE	1.03	1.00	1.00	1.00
Maximum SE	1.13	1.00	1.00	1.00

(b) Model A2

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
Average SE	1.39	1.27	1.26	1.00
Maximum SE	3.61	1.97	1.80	1.00

Under Model A1, since both rerandomization and the FSM are able to adequately balance the means of the original covariates, they lead to lower SE (hence, higher efficiency) than CRD. Across randomizations, the worst case SE under RR 0.01, RR 0.001, and the FSM are 13% smaller than under CRD. Under Model A1, the FSM has similar model-based SE as the two rerandomization methods. However, under Model A2, the FSM uniformly outperforms the other three designs, with a 26% reduction in average SE and an 80% reduction in maximum SE than RR 0.001. This improvement in efficiency can be attributed to the balance achieved by the FSM on the main covariates and their squares and pairwise products. In sum, when the model assumed at the design stage is correct and is used at the analysis stage, the FSM is as efficient as the two rerandomizations for estimating the treatment effect. However, when the model assumed at the design stage is misspecified and later corrected by augmenting transformations of the covariates (e.g., squares and pairwise products), the FSM is considerably more efficient and robust than the other designs.

Under the randomization-based approach, we compare the standard errors of the difference-in-means statistic under each design. Following Hainmueller (2012), the potential outcomes are generated using the models: $Y(1) = X_1+X_2+X_3-X_4+X_5+X_6+\eta$, $Y(2) = Y(1)$ (Model B1) and $Y(1) = (X_1 + X_2 + X_5)^2 + \eta$, $Y(2) = Y(1)$ (Model B2), where $\eta \sim \mathcal{N}(0, 1)$. Both generative models satisfy the sharp-null hypothesis of zero treatment effect for every unit and hence, $\text{SATE}_{2,1} = 0$. Conditional on these potential outcomes, $\text{SATE}_{2,1}$ is estimated

under each design using the standard difference-in-means estimator. The corresponding randomization-based SE of this estimator is obtained by generating 800 randomizations of the design and computing the standard deviation of the difference-in-means estimator across these 800 randomizations. Table A5 shows the randomization-based SE of the difference-in-means statistic for $\text{SATE}_{2,1}$ under each model.

Table A5: Randomization-based standard errors relative to the FSM. The standard error for the FSM is 0.2 under Model B1 (linear model on the main covariates) and 0.43 under Model B2 (linear model on the main covariates and their second-order transformations). Especially under Model B2, the FSM is considerably more efficient than both CRD and RR.

(a) Model B1

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
SE	2.72	1.26	1.08	1

(b) Model B2

	Designs			
	CRD	RR 0.01	RR 0.001	FSM
SE	5.69	4.56	4.47	1

Under Model B1, the potential outcomes depend linearly on the covariates and therefore balancing the means of the covariates improves efficiency. This is reflected in Table A5 as the FSM has the smallest SE, closely followed by RR 0.001. Under Model B2, the potential outcomes depend linearly on the squares and pairwise products of the covariates. By better balancing these transformations, the FSM yields a considerably smaller SE than the other designs. In particular, under Model B2, the SE under the FSM is 67% smaller than the SE under RR 0.001. Therefore, as in the model-based approach, in the randomization-based approach the FSM exhibits comparable efficiency to rerandomization under correct-specification of the outcome model and considerable robustness under model misspecification.

H.4 Comparison with the global D-optimal design

In this section, we compare the performance of the FSM with the global D-optimal design (or simply, the D-optimal design), as defined in Section H.4. Obtaining the exact D-optimal assignment is an NP-hard problem in general, so we consider two alternatives. First, we randomly sample a large number (20000) of assignment vectors from the space of all possible

assignments and obtain the D-optimal assignment among them. Due to random sampling, this assignment is expected to have similar properties (e.g., balance) as the exact D-optimal assignment. Second, we consider a random subsample of 20 units from the original sample of 120 units. For this subsample, we compare FSM with the D-optimal assignment. In this case, both the designs assign units into two groups of size 10 each.

Tables A6 and A7 display the average ASMD values for the original covariates, as well as their squares and interactions, respectively, under the first scenario. Correspondingly, Tables A8 and A9 present these ASMD values under the second scenario.

Table A6: ASMD of the original covariates under the D-optimal design (D-opt), and the average ASMD of the original covariates under the FSM. The ASMD for the D-optimal design is approximated based on 20000 randomizations.

Covariates	Designs	
	D-opt	FSM
X_1	0.031	0.029
X_2	0.008	0.025
X_3	0.020	0.042
X_4	0.004	0.029
X_5	0.041	0.029
X_6	0.033	0.034
Mean	0.023	0.031

Table A7: ASMD of the squares and pairwise products of the covariates under the D-optimal design (D-opt), and the average ASMD of the same transformations under the FSM. The ASMD for the D-optimal design is approximated based on 20000 randomizations.

Covariate transformations	Designs	
	CRD	RR 0.01
X_1X_2	0.029	0.041
X_1X_2	0.038	0.041
X_1X_2	0.206	0.024
X_1X_2	0.074	0.035
X_1X_2	0.223	0.030
X_1X_2	0.057	0.051
X_1X_2	0.090	0.027
X_1X_2	0.027	0.030
X_1X_2	0.075	0.026
X_1X_2	0.329	0.032
X_1X_2	0.147	0.096
X_1X_2	0.087	0.035
X_1X_2	0.064	0.037
X_1X_2	0.091	0.027
X_1X_2	0.036	0.024
X_1^2	0.029	0.031
X_2^2	0.085	0.038
X_3^2	0.110	0.041
X_4^2	0.060	0.053
X_5^2	0.047	0.013
Mean	0.095	0.037

Table A8: ASMD of the original covariates in the sampled dataset under the D-optimal design (D-opt), and the average ASMD of the original covariates under the FSM.

Covariates	Designs	
	D-opt	FSM
X_1	0.022	0.191
X_2	0.071	0.142
X_3	0.036	0.213
X_4	0.016	0.147
X_5	0.054	0.194
X_6	0.000	0.051
Mean	0.033	0.156

Table A9: ASMD of the squares and pairwise products of the covariates in the sampled dataset under the D-optimal design (D-opt), and the average ASMD of the same transformations under the FSM.

Covariate transformations	Designs	
	CRD	RR 0.01
X_1X_2	0.825	0.353
X_1X_2	1.231	0.210
X_1X_2	0.484	0.162
X_1X_2	0.588	0.388
X_1X_2	0.727	0.230
X_1X_2	1.526	0.248
X_1X_2	0.765	0.095
X_1X_2	0.625	0.363
X_1X_2	0.477	0.264
X_1X_2	0.638	0.269
X_1X_2	0.740	0.392
X_1X_2	0.440	0.263
X_1X_2	0.559	0.404
X_1X_2	0.609	0.147
X_1X_2	0.238	0.063
X_1^2	0.952	0.200
X_2^2	0.116	0.365
X_3^2	0.833	0.313
X_4^2	0.566	0.167
X_5^2	0.019	0.233
Mean	0.648	0.256

From the above tables we observe that, on an average, the D-optimal design produces better balance on the main covariates. This observation is consistent with Proposition A5, which shows that with a single covariate, the D-optimal design aims to exactly balance its mean across the two groups. However, akin to randomization, it produces worse balance on the second-order transformations of the covariates.

H.5 Additional results from the simulation study

Table A10: Averages of the ASMD of the original covariates across 800 randomizations.

Covariates	Designs			
	CRD	RR 0.01	RR 0.001	FSM
X_1	0.162	0.051	0.035	0.029
X_2	0.156	0.048	0.033	0.025
X_3	0.158	0.049	0.033	0.042
X_4	0.150	0.049	0.034	0.029
X_5	0.140	0.052	0.034	0.029
X_6	0.141	0.052	0.036	0.034
Mean	0.151	0.050	0.034	0.031

Table A11: Averages of the ASMD of squares, pairwise products, and other transformations of the covariates across 800 randomizations.

Covariate transformations	Designs			
	CRD	RR 0.01	RR 0.001	FSM
X_1X_2	0.144	0.153	0.148	0.041
X_1X_3	0.144	0.140	0.137	0.041
X_1X_4	0.141	0.148	0.147	0.023
X_1X_5	0.150	0.135	0.134	0.035
X_1X_6	0.152	0.109	0.101	0.030
X_2X_3	0.147	0.147	0.146	0.051
X_2X_4	0.140	0.155	0.150	0.027
X_2X_5	0.147	0.143	0.136	0.030
X_2X_6	0.152	0.115	0.104	0.026
X_3X_4	0.141	0.143	0.152	0.032
X_3X_5	0.149	0.140	0.139	0.096
X_3X_6	0.148	0.099	0.091	0.035
X_4X_5	0.148	0.132	0.130	0.037
X_4X_6	0.152	0.100	0.095	0.027
X_5X_6	0.146	0.095	0.094	0.024
X_1^2	0.140	0.145	0.143	0.031
X_2^2	0.151	0.155	0.150	0.038
X_3^2	0.144	0.136	0.132	0.041
X_4^2	0.143	0.145	0.147	0.053
X_5^2	0.142	0.073	0.067	0.013
Mean	0.146	0.130	0.127	0.037
$X_5^{1.5}$	0.141	0.060	0.048	0.018
X_2^3	0.155	0.090	0.081	0.071
X_4^4	0.140	0.143	0.147	0.072
$\frac{1}{4+X_3}$	0.157	0.073	0.064	0.050
Mean	0.148	0.092	0.085	0.053

Figure A2: Boxplot of the distribution of $||\underline{\boldsymbol{S}}_1 - \underline{\boldsymbol{S}}_2||_F$ across 800 randomizations, where $\underline{\boldsymbol{S}}_g$ is the sample covariance matrix of the covariates in treatment group $g \in \{1, 2\}$.

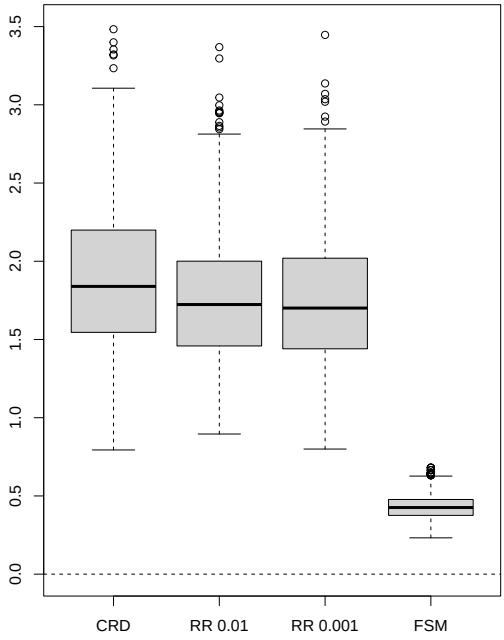


Table A12: Averages of the ASMD of the original covariates across 800 randomizations under the FSM with differences choices of ϵ.

Covariates	Choice of ϵ			
	0.1	0.01	0.001	0.0001
X_1	0.032	0.030	0.030	0.030
X_2	0.029	0.026	0.026	0.026
X_3	0.041	0.043	0.043	0.043
X_4	0.026	0.028	0.028	0.028
X_5	0.029	0.031	0.031	0.031
X_6	0.034	0.034	0.034	0.034
Mean	0.032	0.032	0.032	0.032

Table A13: Averages of the ASMD of squares and pairwise products of the covariates across 800 randomizations under the FSM with different choices of ϵ.

Covariate transformations	Choice of ϵ			
	0.1	0.01	0.001	0.0001
X_1X_2	0.044	0.038	0.038	0.038
X_1X_3	0.040	0.041	0.041	0.041
X_1X_4	0.028	0.025	0.025	0.025
X_1X_5	0.039	0.037	0.037	0.037
X_1X_6	0.031	0.030	0.030	0.030
X_2X_3	0.045	0.048	0.048	0.048
X_2X_4	0.029	0.026	0.026	0.026
X_2X_5	0.040	0.029	0.029	0.029
X_2X_6	0.028	0.026	0.026	0.026
X_3X_4	0.038	0.033	0.033	0.033
X_3X_5	0.091	0.097	0.097	0.097
X_3X_6	0.028	0.033	0.033	0.033
X_4X_5	0.046	0.038	0.038	0.038
X_4X_6	0.026	0.027	0.027	0.027
X_5X_6	0.024	0.027	0.027	0.027
X_1^2	0.031	0.032	0.032	0.032
X_2^2	0.038	0.036	0.036	0.036
X_3^2	0.040	0.040	0.040	0.040
X_4^2	0.052	0.052	0.052	0.052
X_5^2	0.011	0.014	0.014	0.014
Mean	0.037	0.036	0.036	0.036

I Additional results from the Health Insurance Experiment

Table A14: Average ASMD of the main covariates between treatment groups 1 and 2 across 400 randomizations.

Covariates	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
X_1 : Log family size	0.052	0.039	0.038	0.012
X_2 : Log family income	0.052	0.040	0.043	0.010
X_3 : Max hourly wage	0.051	0.042	0.047	0.017
X_4 : Adult med visits	0.049	0.043	0.041	0.014
X_5 : Kid med visits	0.048	0.039	0.040	0.010
X_6 : Female	0.047	0.039	0.040	0.010
X_7 : Age 0 to 5	0.053	0.038	0.039	0.010
X_8 : Age 6 to 17	0.051	0.041	0.039	0.011
X_9 : Age 18 to 44	0.053	0.038	0.040	0.010
X_{10} : Male HS Grad	0.051	0.038	0.041	0.006
X_{11} : Male more than HS	0.048	0.037	0.041	0.006
X_{12} : Insured	0.049	0.040	0.038	0.010
X_{13} : Excellent health	0.052	0.040	0.037	0.009
X_{14} : Good health	0.053	0.038	0.037	0.010
X_{15} : Family income mis	0.052	0.038	0.041	0.011
X_{16} : Max hourly wage mis	0.051	0.038	0.041	0.013
X_{17} : Adult med visits mis	0.054	0.040	0.040	0.011
X_{18} : Kid med visits mis	0.057	0.041	0.039	0.011
X_{19} : Education male mis	0.048	0.038	0.041	0.008
X_{20} : Insured mis	0.048	0.039	0.038	0.011
Mean	0.051	0.039	0.040	0.011

Table A15: Averages of the ASMD between each pair of treatment groups across the original covariates and across 400 randomizations.

Treatment group	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
$1, 2$	0.051	0.039	0.040	0.011
$1, 3$	0.055	0.041	0.043	0.011
$1, 4$	0.049	0.038	0.039	0.010
$2, 3$	0.056	0.043	0.045	0.012
$2, 4$	0.053	0.040	0.041	0.010
$3, 4$	0.056	0.042	0.044	0.012
Mean	0.053	0.040	0.042	0.011

Table A16: Averages of the ASMD of the squares and pairwise products of the (demeaned) covariates X_1,..., X_5 between treatment groups 1 and 2 across 400 randomizations.

Covariates	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
X_1X_2	0.053	0.039	0.041	0.020
X_1X_3	0.053	0.047	0.046	0.027
X_1X_4	0.054	0.045	0.045	0.020
X_1X_5	0.049	0.040	0.041	0.013
X_2X_3	0.054	0.049	0.053	0.038
X_2X_4	0.050	0.045	0.048	0.017
X_2X_5	0.052	0.039	0.039	0.015
X_3X_4	0.054	0.043	0.045	0.022
X_3X_5	0.050	0.042	0.046	0.022
X_4X_5	0.054	0.044	0.045	0.015
X_1^2	0.053	0.041	0.040	0.026
X_2^2	0.051	0.041	0.042	0.015
X_3^2	0.057	0.055	0.058	0.026
X_4^2	0.053	0.053	0.053	0.012
X_5^2	0.051	0.043	0.044	0.004
Mean	0.053	0.044	0.046	0.019

Table A17: Averages of the ASMD between each pair of treatment groups across the squares and pairwise products of the (demeaned) covariates X_1,..., X_5 and across 400 randomizations.

Treatment group	Designs			
	CRD	RR 0.01	RR 0.001	FSM
$1, 2$	0.053	0.044	0.046	0.019
$1, 3$	0.056	0.046	0.048	0.020
$1, 4$	0.051	0.043	0.044	0.017
$2, 3$	0.058	0.049	0.049	0.021
$2, 4$	0.054	0.046	0.046	0.018
$3, 4$	0.058	0.048	0.049	0.023
Mean	0.055	0.046	0.047	0.020

Figure A3: Distributions of ASMD of all cubes and three-way interactions of the non-binary covariates across randomizations.

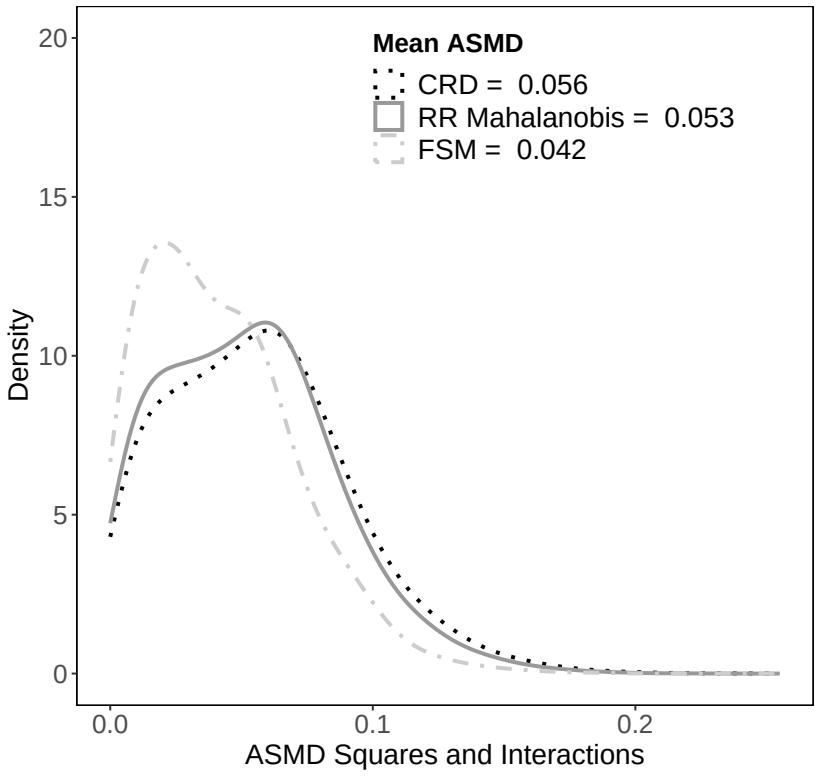
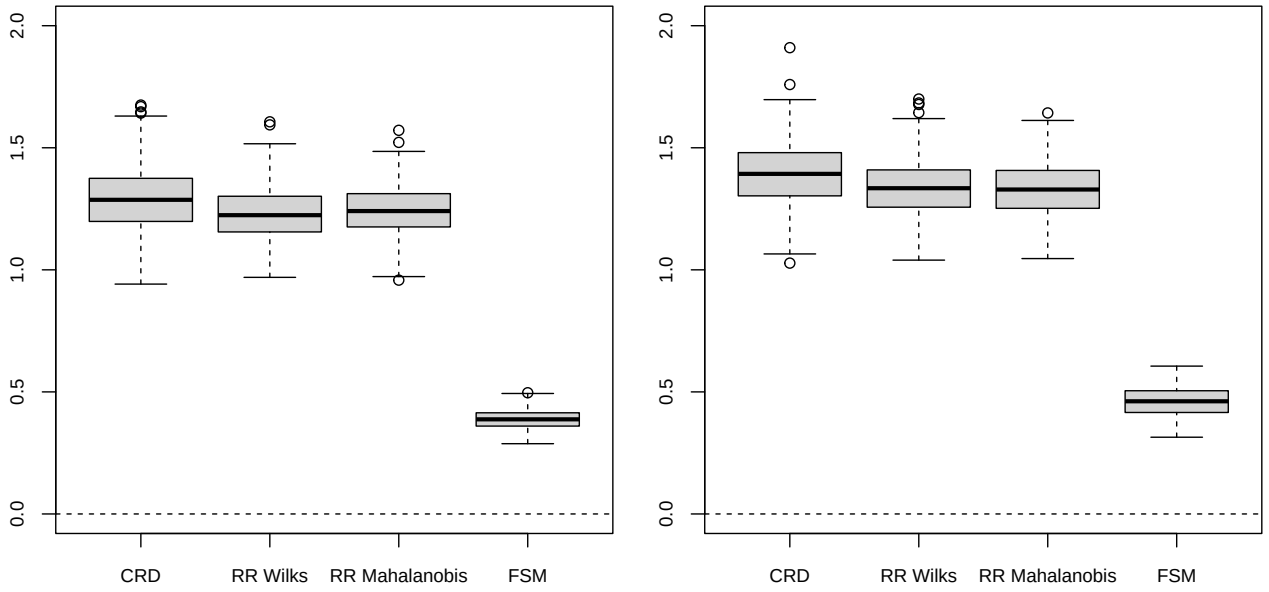
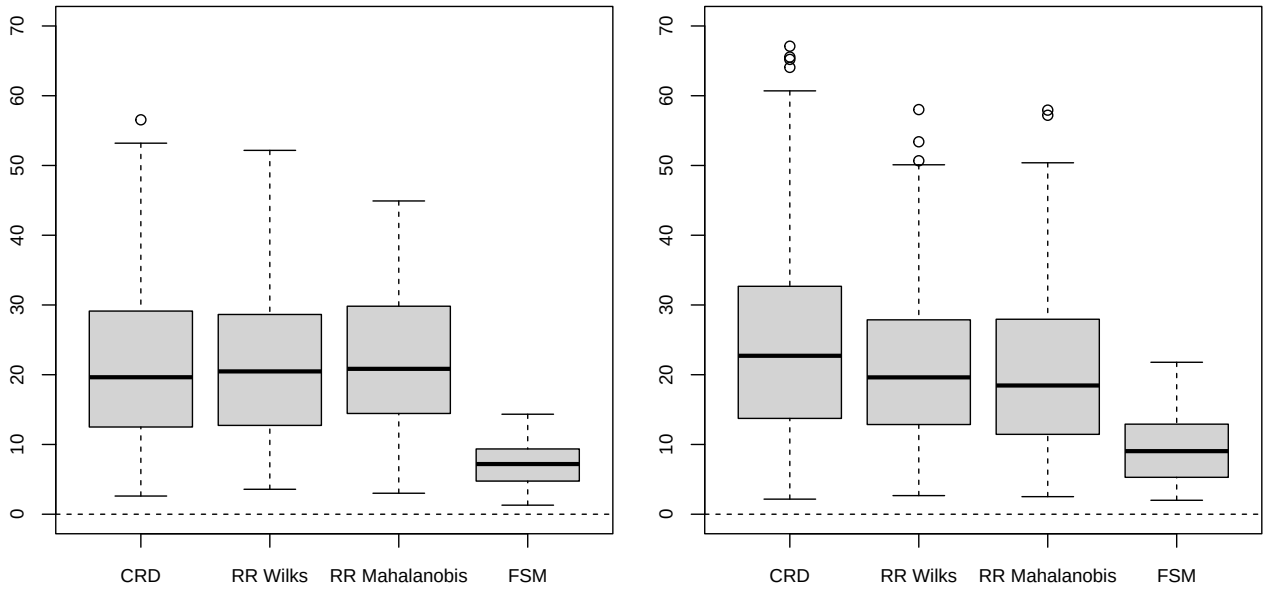


Figure A4: Distributions of discrepancies of the correlation matrices of the covariates in the treatment groups of the HIE data across randomizations. The discrepancies are measured by $||\underline{\boldsymbol{R}}_g - \underline{\boldsymbol{R}}_{g'}||_F$, where $\underline{\boldsymbol{R}}_g$ is the sample correlation matrix of the covariates in treatment group g and $||\cdot||_F$ is the Frobenius norm. The FSM systematically produces lower discrepancies than the other methods, exhibiting substantially improved balance on the correlations of the covariates.



(a) $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$ (b) $||\underline{\boldsymbol{R}}_3 - \underline{\boldsymbol{R}}_4||_F$

Figure A5: Boxplot of the distribution of $||\underline{\boldsymbol{S}}_g - \underline{\boldsymbol{S}}_{g'}||_F$ across 400 randomizations, where $\underline{\boldsymbol{S}}_g$ is the sample covariance matrix of the covariates in treatment group $g \in \{1, 2, 3, 4\}$.



(a) $||\underline{\boldsymbol{S}}_1 - \underline{\boldsymbol{S}}_2||_F$ (b) $||\underline{\boldsymbol{S}}_3 - \underline{\boldsymbol{S}}_4||_F$

We now compare the efficiency of the designs in the randomization-based approach with four additional potential outcome models given below.

- Model B3: $Y(3) = 5 - 3X_1 + X_2 + X_3 - 0.2X_4 + 0.8X_5 + \eta$, $Y(3) = Y(2)$.
- Model B4: $Y(3) = 5 - 2X_1^2 + 0.5X_3^2 + 0.5X_5^2 + 5X_1X_2 - 0.8X_3X_5 + \eta$, $Y(3) = Y(2)$.
- Model B5: $Y(3) = 10 + 8X_1X_2 + 3X_2X_5 - 0.5X_3X_5 + \eta$, $Y(3) = Y(2)$.
- Model B6: $Y(3) = 0.8X_1X_2 - 3X_3^2 + \frac{1}{1+X_4} - 4X_1^3 + \eta$

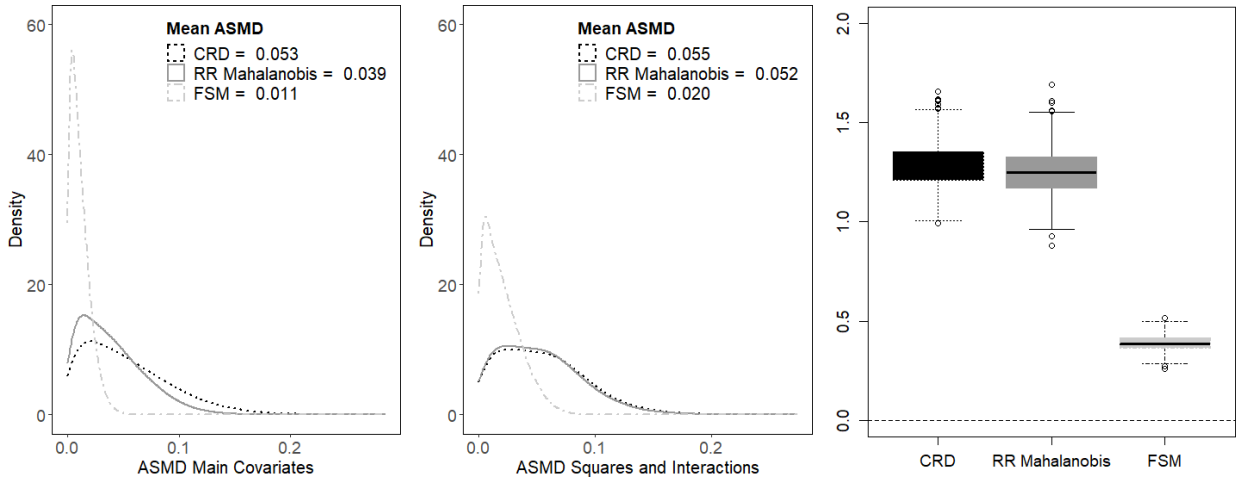
For each model, the error term $\eta \sim \mathcal{N}(0, 1.5^2)$. Under each design, $\text{SATE}_{3,2}$ is estimated using the standard difference-in-means estimator and the corresponding randomization-based SE is obtained by generating 400 randomizations and computing the standard deviation of the estimator across these 400 randomizations. The average randomization-based standard errors (across 500 simulations) are presented in Table A18.

Table A18: Average randomization-based standard errors relative to the FSM under Models B3, B4, B5, B6

	Designs			
	CRD	RR Wilks	RR Mahalanobis	FSM
Model B3	2.36	1.80	1.90	1
Model B4	2.14	1.75	1.81	1
Model B5	2.99	2.40	2.44	1
Model B6	1.61	1.42	1.44	1

We finish this section by evaluating and comparing the covariate balance on the main covariates and the second-order transformations, for CRD, RR, and the FSM, where RR uses the Mahalanobis distance based on the main covariates only and accepts 0.1% of the assignments.

Figure A6: Distributions of absolute standardized mean differences (ASMD) of the main covariates (panel (a)) and the squares and pairwise products of the scaled covariates (panel (b)) across randomizations. For each plot, the legend presents the average ASMD across simulations for the four methods. Panel (c) shows distributions of discrepancies between the correlation matrices of the covariates in treatment groups 1 and 2 (as measured by the Frobenius norm, $||\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2||_F$). In terms of the main covariates, second-order transformations, and correlation matrices, the FSM substantially outperforms CRD and RR.



(a) Main covariates (b) Squares and pairwise products (c) Frobenius norm

Figure A6 shows a pattern of performances of the designs akin to those illustrated in Section 6. The FSM, outperforms CRD and RR in terms of balancing both the main covariates and their second-order transformations. As compared to the previous version, this version of RR reduces the average imbalance on the main covariates, while increasing the average imbalance on the second-order transformations. This behavior aligns with our expectations, since this version of RR specifically targets balance on the main covariates, not on their second order transformations.

J Additional figures from the case studies

Figure A7: Distributions of the absolute standardized mean differences of the main covariates and their squares and interactions, and the Frobenius norms of $\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2$ under complete randomization, rerandomization, and the FSM, for the five studies: (1) Angrist, (2) Blattman, (3) Durocher, (4) Finkelstein, (5) Lalonde.

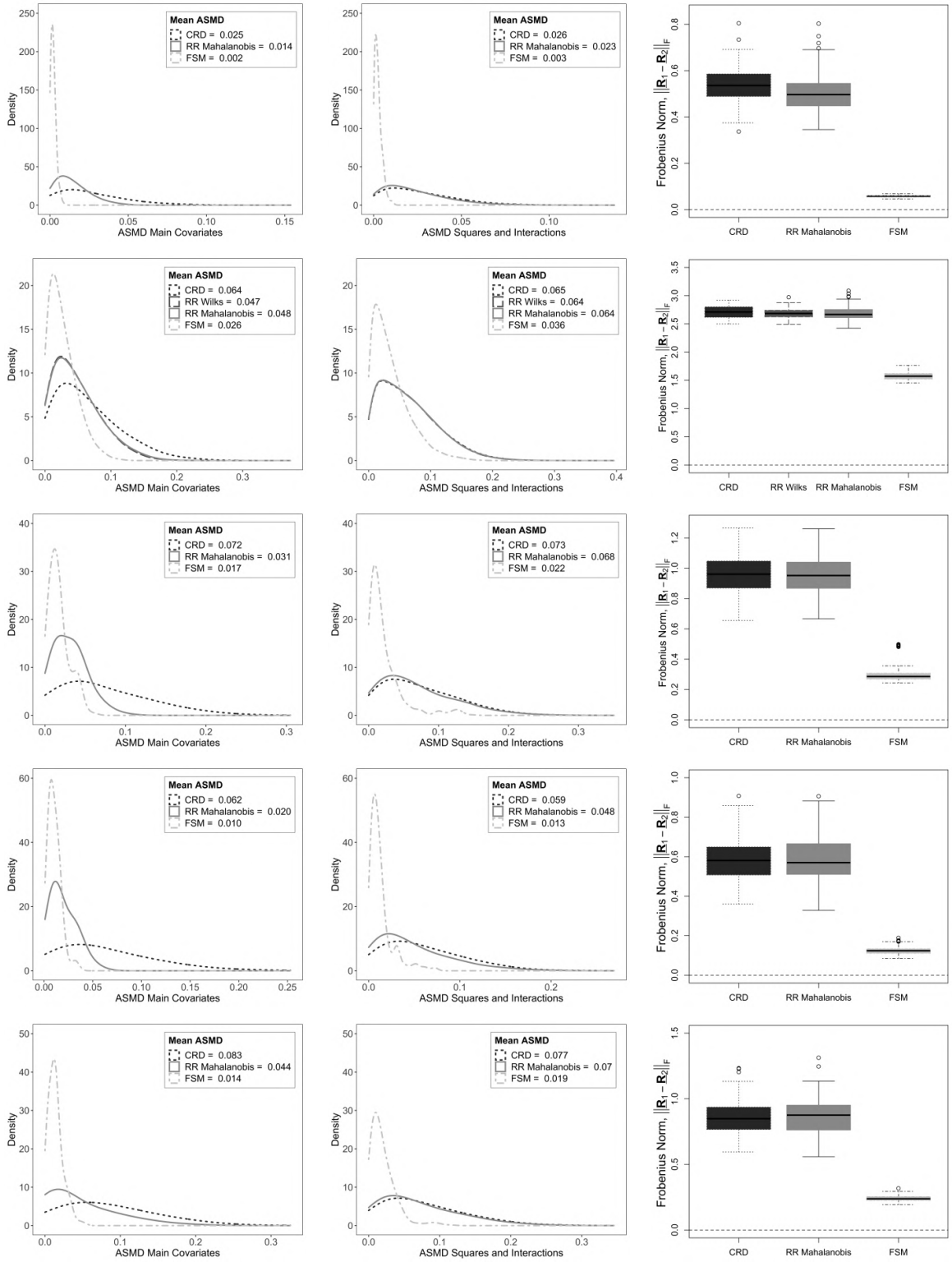
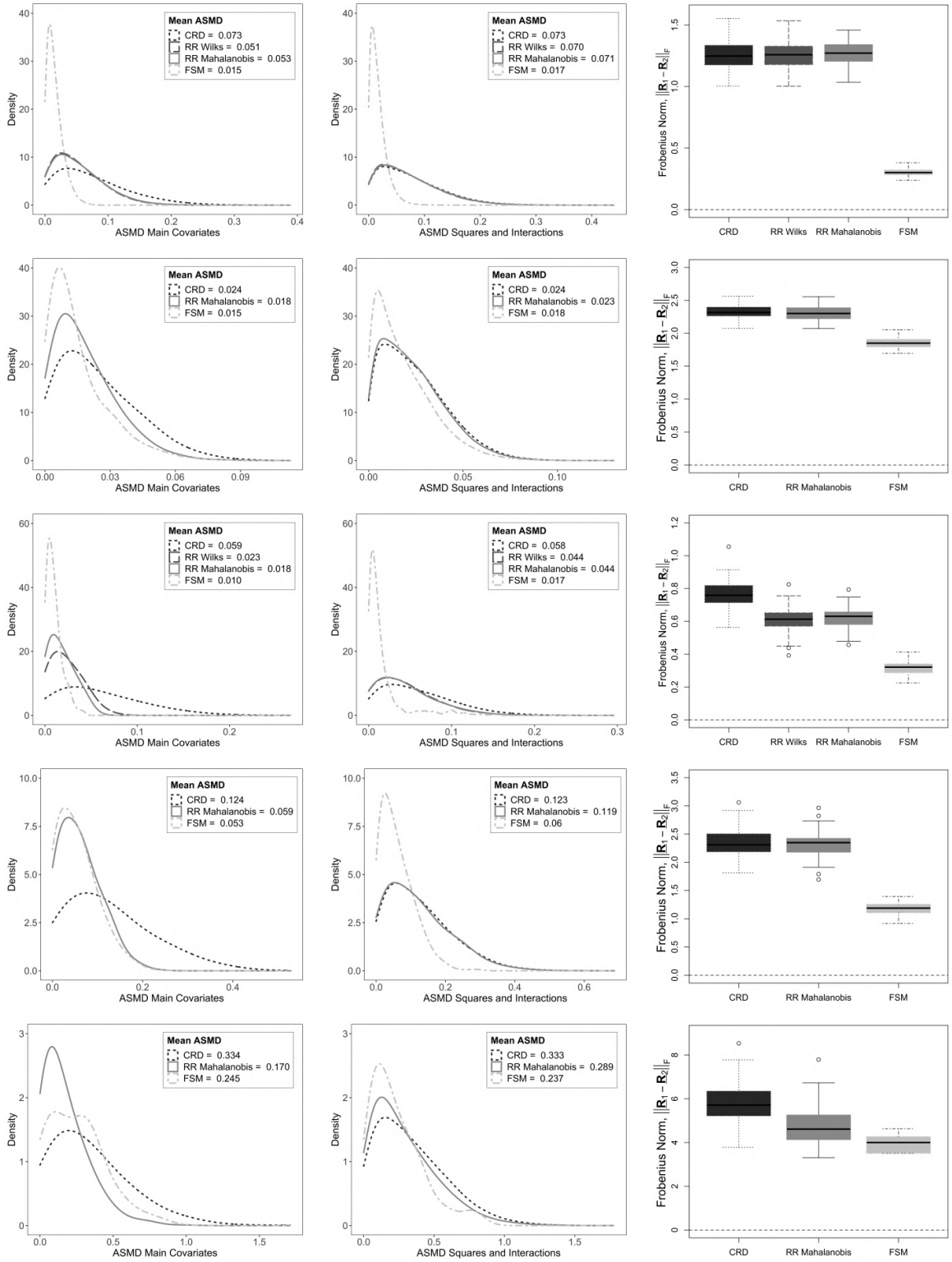


Figure A8: Distributions of the absolute standardized mean differences of the main covariates and their squares and interactions, and the Frobenius norms of $\underline{\boldsymbol{R}}_1 - \underline{\boldsymbol{R}}_2$ under complete randomization, rerandomization, and the FSM, for the five studies: (6) Ambler, (7) Crepon, (8) Dupas, (9) Karlan, (10) Wantchekon.



References

Aronow, P. M. and Samii, C. (2013), "Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities," *Survey Methodology*, 39, 231–241.

Chattopadhyay, A., Morris, C. N., and Zubizarreta, J. R. (2021), "Randomized and Balanced Allocation of Units into Treatment Groups Using the Finite Selection Model for R," *arXiv preprint arXiv:2105.02393*.

Hainmueller, J. (2012), "Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Morris, C. (1979), "A finite selection model for experimental design of the health insurance study," *Journal of Econometrics*, 11, 43–61.

— (1983), "Sequentially controlled Markovian random sampling (SCOMARS)," *Institute of Mathematical Statistics Bulletin*, 12, 237.

Morris, C. and Hill, J. (2000), "The health insurance experiment: design using the finite selection model," *Public Policy and Statistics: Case Studies from RAND*, Springer Science & Business Media, 29–53.

Mukerjee, R., Dasgupta, T., and Rubin, D. B. (2018), "Using standard tools from finite population sampling to improve causal inference for complex experiments," *Journal of the American Statistical Association*, 113, 868–881.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.

Stuart, E. A. (2010), "Matching methods for causal inference: a review and a look forward," *Statistical Science*, 25, 1–21.